Analysis of diurnal gene regulation and metabolic diversity in Synechocystis sp. PCC 6803 and other phototrophic cyanobacteria

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der

Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

vorgelegt von

Diplom-Bioinformatiker
Johannes Christian Beck

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin:
Prof. Dr. Bernhard Grimm

Gutachter/innen:

1. Prof. Dr. Hanspeter Herzel
2. Jun.-Prof. Dr. Ilka Maria Axmann
3. Dr. Ralf Steuer

Tag der mündlichen Prüfung: 17. April 2018

Acknowledgements

With my first words, I want to use the opportunity to express my gratitude to all the people who supported me and my work on various levels, be it professional or personal. Without their help, this thesis would not have been finished.

First and foremost, I want to express my deepest gratitude to my beloved wife Evelin, who not only gave me a lovely son, but inspired me throughout the years and always stood by my side when I was tumbling. Without her love, assistance, and unfailing support this work could not have been accomplished.

I want to thank Professor Dr. Hanspeter Herzel for his support and for giving me the opportunity to carry out my research at the Institute for Theoretical Biology.

My gratitude also goes to Dr. Ralf Steuer, for his continuing support of my work and for inspiring me with his comprehensive knowledge and his interdisciplinary points of view. Thanks for all the stimulating discussions and the nonsense chit-chat.

Special thanks go to Jun.-Professor Dr. Ilka Maria Axmann who warmly invited me into her group and gave me the opportunity to work on the circadian clock. I will always remember her passion for science as well as her passion for sociality.

Very special gratitudes go out to the colleagues turned friends Dr. Henning Knoop for his seemingly endless knowledge about the bacterial metabolism, and Dr. Bharath Ananthasubramaniam for his support with all sorts of technical difficulties.

I also want to thank all the people from Dr. Ilka Axmann's group: Beate Heilmann, Anika Wiegard, Adrian Kölsch, and Nicolas Schmelling for always welcoming me into the lab, but especially Dr. Anne Rediger for her meticulous work in conducting the time-series experiments and Dr. Stefanie Hertel for her knowledge of the cyanobacterial clock and the interesting discussions during endless nights of sampling in the lab.

Moreover, I want to thank the people from the ITB, especially Marjan Faizi, Rukeia El-Athman, Mónica Abreu, Nikolai Genov, and Janek Grzegorzewski who every day made it easier to go to work and to the "Mensa".

And last, but by no means least, I want to send my deepest thanks to my friends Manuela and Uwe, Juliane and Sebastian, as well as my parents and siblings for their continuous support, open arms, and endless love.

Abstract

Cyanobacteria are an ancient, highly diverse phylum of prokaryotes, populating virtually all habitats on the surface of the earth. Most organisms of this group are capable of a photoautotrophic lifestyle and harvest energy by capturing photons emitted from the sun to convert atmospheric carbon dioxide into essential organic compounds. This makes them one of the prime producers for the global food chain. Due to the rhythmic alternation of day and night, cyanobacteria constantly have to cope with variations in the availability of their primary energy source. Most strains therefore harbor a circadian clock mechanism consisting of three proteins KaiA, KaiB, and KaiC, whose biochemical interactions result in a self-sustained and temperature-compensated phosphorylation cycle with a period of approximately 24 hours. Understanding circadian gene expression, facilitated by this mechanism, is a major objective of this thesis. Three time-series experiments with distinct patterns of illumination were conducted in the model organism *Synechocystis* sp. PCC 6803. Despite its possession of *kaiA* and three copies each of *kaiB* and *kaiC*, I did not detect any self-sustained oscillations in this organism. However, genes were expressed in a tight diurnal schedule corresponding to the photoautotrophic lifestyle, in oscillating light-dark conditions. In addition, I observed a novel phenomenon of diurnal accumulation of ribosomal RNAs during dark periods. This finding challenges common assumptions on the amount of ribosomal RNAs and I can only speculate about the association with growth in rhythmic environments.

Due to their high growth rates and undemanding nature regarding nutrients, water salinity, and temperature, cyanobacteria emerged as a viable option for sustainable production of various chemical commodities, including biofuels, biopolymers, or plain biomass. For cost-efficient production rates, however, optimization of growth conditions and genetic modification of intracellular biochemical reactions are inevitable, both requiring comprehensive knowledge of the cyanobacterial metabolism. To address this issue, I compared the genomes of multiple cyanobacterial strains through an exhaustive analysis of orthologous genes, which allowed reasonable guesses for the size of the full complement of genes as well as the set of highly conserved core genes. Examining the diversity of the central carbon and storage metabolism, I observed that parts of the network are highly conserved, while other pathways are adjusted to the organism's modes of life. Systematic analysis of genes shared by similar subsets of organisms indicates high rates of functional relationship in such co-occurring genes. Going one step further, I designed a novel approach to identify modules of co-occurring genes, which exhibit a high degree of functional coherence and reveal known but also previously unknown functional relationships. Complementing the precomputed modules, I developed the SimilarityViewer, a graphical toolbox that facilitates further analysis of co-occurrence with respect to specific cyanobacterial genes of interest. Simulations of automatically generated metabolic reconstructions revealed the biosynthetic capacities of individual cyanobacterial strains, which will assist future research addressing metabolic engineering of cyanobacteria.

Zusammenfassung

Cyanobakterien bilden einen alten, weitverzweigten Stamm von Bakterien und finden sich in nahezu allen Biotopen auf der Oberfläche der Erde. Die meisten Organismen dieser Gruppe leben photoautotroph und beziehen ihre Energie aus dem Sonnenlicht, um Kohlenstoffdioxid aus der Atmosphäre in essentielle organische Bausteine umzuwandeln. Damit gehören sie zu den wichtigsten Produzenten der weltweiten Nahrungskette. Aufgrund des regelmäßigen Wechsels von Tag und Nacht müssen sich Cyanobakterien ständig auf veränderte Lichtverhältnisse einstellen. Die meisten Unterarten besitzen dafür eine innere Uhr, welche aus den drei Proteinen KaiA, KaiB und KaiC besteht, deren biochemische Interaktionen zu einem selbsterhaltenden und temperaturunabhängigen 24-stündigen Rhythmus von Phosphorylierung und Dephosphorylierung führen. Um die von diesem Mechanismus ausgelöste circadiane Genexpression besser zu verstehen, wurden drei Zeitserienexperimente mit unterschiedlicher Beleuchtung in dem Modellorganismus *Synechocystis* sp. PCC 6803 durchgeführt. Obwohl dieses Bakterium neben *kaiA* auch jeweils drei Genkopien von *kaiB* und *kaiC* besitzt, konnte ich keine selbsterhaltenden Rhythmen entdecken. Gleichwohl werden die Gene in einer Tag-Nacht-Umgebung nach einem genauen Zeitplan aktiviert. Des Weiteren konnte ich einen starken Anstieg der ribosomalen RNA in der Dunkelheit beobachten, wobei dies dem allgemeinen Wissen über die zelluläre Konzentration von RNA widerspricht.

In den letzten Jahren haben Cyanobakterien aufgrund ihrer hohen Wachstumsraten und geringen Anforderungen an Nährstoffe, Wasserqualität und Temperatur hohes Interesse für die nachhaltige Erzeugung von Biokraftstoff, Biokunststoff oder Biomasse hervorgerufen. Für einen wettbewerbsfähigen, industriellen Einsatz werden allerdings optimale Wachstumsbedingungen und genetische Optimierungen benötigt, was umfangreiches Wissen über den Metabolismus der Cyanobakterien voraussetzt. Dafür habe ich die Orthologien zwischen zahlreichen Cyanobakterien untersucht, was unter anderem eine gute Abschätzung der Anzahl aller verfügbaren sowie aller konservierten Gene ermöglicht. Zusätzlich konnte ich beobachten, dass ein Teil des zentralen Kohlenstoffmetabolismus speziesübergreifend konserviert ist, während andere Teile stark an die Lebensweise der Organismen angepasst sind. Eine systematische Analyse von Genen, die gemeinsam in denselben Arten vorkommen, zeigt, dass diese Gene oft auch eine gemeinsame Funktion haben. Ich habe daher eine Methode entworfen, um Gruppen von gemeinsam vorkommenden Genen zu identifizieren. Ich konnte zeigen, dass diese häufig an zusammengehörigen biologischen Prozessen beteiligt sind und damit bekannte, aber auch unbekannte Beziehungen aufdecken. Zusätzlich zu den in der Arbeit diskutierten Modulen habe ich den SimilarityViewer entwickelt, ein grafisches Computerprogramm für die Identifizierung von gemeinsam vorkommenden Partnern für jedes beliebige Gen. Außerdem habe ich für alle Organismen automatische Rekonstruktionen des Stoffwechsels erstellt und konnte zeigen, dass diese die Synthese von gewünschten Stoffen gut vorhersagen, was hilfreich für zukünftige Forschung am Metabolismus von Cyanobakterien sein wird.

Contents

Chapter 1.

Introduction

Cyanobacteria, also known as Cyanophyta or sometimes incorrectly called blue-green algae, are unicellular, gram-negative bacteria mostly capable of oxygenic photosynthesis. During this complex process electromagnetic energy is harvested from photons - naturally emitted from the sun - and converted to chemical energy stored in energy-rich molecules like adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide phosphate (NADPH). This transformation requires a series of chemical steps and multiple enzymes. In short, photons are captured by photosystem II and the energy is transferred to electrons; the excited electrons are transported via the electron transport chain to photosystem I; the redox potential of the electron is raised again by another transfer of photon energy; finally the electrons are used to reduce NADP+ to NADPH. Along that chain of reactions, protons are pumped out of the thylakoid lumen to create an electrochemical gradient at the thylakoid membrane, which is used to synthesize ATP. Electrons in photosystem II are gained by splitting water, creating protons and molecular oxygen as waste products (Peschek et al. 2011). This process is carried out in cyanobacteria as well as the chloroplasts of plants and algae and provides the energy needed for conversion of atmospheric carbon dioxide to organic matter through the Calvin-Benson cycle. Roughly one percent of the sun's light hitting the earth is harnessed through photosynthesis (Pisciotta et al. 2010) and utilized for a global gross production of over 150 petagrams of carbon per year (Welp et al. 2011), making phototrophic growth the most important process for the biological life on earth. Around 30% of world-wide carbon production can be attributed to photosynthetic prokaryotes alone (Pisciotta et al. 2010).

1.1. Cyanobacteria: from past to present

Cyanobacteria are a very ancient phylum of bacteria. Life emerged roughly 3.8 billion years ago (Cavalier-Smith 2002), and first forms of cyanobacteria might date back more than 3.2 billion years (Whitton 2012; Kaufman 2014). These protocyanobacteria most likely already possessed the two photosystems common in todays cyanobacteria and plants, but were only capable of anoxygenic photosynthesis using hydrogen, hydrogen peroxide, or hydrogen sulfide as electron donor instead of water (Blankenship 2001; Sousa et al. 2013). At that time, the dioxygen concentration of the atmosphere was at around 0.001% of what it is today (Sessions et al. 2009; Flores and Herrero 2008; Holland 2006). Then, 2.45 billion years ago the oxygen-evolving complex evolved in cyanobacteria allowing them to split water as electron donor, releasing protons and oxygen into the environment (Whitton 2012; Kaufman

2014). This unique event (Flores and Herrero 2008) ultimately led to the Great Oxidation Event (GOE) and transformed geosphere and biosphere of the earth forever. At first, the oxygen was absorbed by oxygen-sinks (e.g. reduced iron and sulfur). When these sinks were filled at around 2.3 billion years ago, oxygen was released to the atmosphere (Kroneck and Torres 2015). Subsequent oxidation of methane and the build-up of the ozone-layer cooled the earth down and made life on the land surface possible (Kasting 2005; Kopp et al. 2005). On the other hand, oxygen was toxic to obligate anaerobic organisms of that time, leading to a mass extinction of these organisms (Rothschild and Lister 2003; Grula 2010). Availability of free oxygen pushed the evolution of aerobic respiration, which is roughly 16 times more efficient than anaerobic fermentation and therefore allowed the development of complex multicellular lifeforms that we know today (Sessions et al. 2009). Cyanobacteria were not only responsible for oxygen evolution but have been the primary producer of organic matter up until the advent of photosynthetic eukaryotic organisms. In a rare endosymbiotic event, one cyanobacterial cell was engulfed by an eukaryotic host cell giving rise to all chloroplasts of modern phototrophic eukaryotes, including land plants and algae (Timmis et al. 2004; Jensen and Leister 2014). The exact time point of this momentous event is highly debated, with estimates ranging from 600 million (Cavalier-Smith 2000), over 1 billion (McFadden 1999), up to 1.5-1.75 billion years ago (Ochoa de Alda et al. 2014) - at least a few hundred million years after the GOE.

Up until today, cyanobacteria play a major role in the global biosphere. The total dry weight of cyanobacterial biomass worldwide is estimated at around 600 million tons, about 1/2000 of the global carbon biomass - most of which is woody tissue (Garcia-Pichel et al. 2003). Moreover, cyanobacteria - especially the most abundant genera *Prochlorococcus* and marine *Synechococcus* - are responsible for roughly 50% of the aquatic and up to 25% of the global primary production of biomass (Kirchman 2012; Goericke and Welschmeyer 1993; Flombaum et al. 2013). Diazotrophic cyanobacteria are able to reduce inert nitrogen N_2 to biologically more accessible compounds such as ammonia. Organisms of the genus Trichodesmium alone are responsible for 43% of the global nitrogen fixation, thus playing a major role in the provision of this vital nutrient (Berman-Frank et al. 2003; Zehr 2011; Latysheva et al. 2012). This property makes cyanobacteria attractive for symbiotic relationships with other organisms. Symbiotic partners include, but are not restricted to all sorts of marine creatures, bacteria, fungi, cycads, algae, plants (Rai et al. 2002; Whitton 2012), and even animals like sponges (Wulff 2006), corals (Lesser et al. 2004), and worms (Carpenter 2002).

In their long history, cyanobacteria adapted to various environments. Today they can be found in virtually all biotopes worldwide, including some of the most hostile environments. Cyanobacteria not only populate nutrient-rich lakes, ponds, rivers, and farm land (Whitton 2012; Sinha and Häder 1996), but also thrive in hot springs (Sompong et al. 2005), cold arctic water ponds (Zakhia et al. 2008), nutrient-depleted oceans (Whitton 2012; Seckbach 2007), and wastewater treatment plants (Martins et al. 2011). Strains adapted to low light conditions grow in deep marine waters (Moore et al. 1998) and caverns (Mulec et al. 2008; Hoffmann 2003). Cyanobacteria can also inhabit non-aqueous environments like glaciers, arctic rocks (Whitton 2012;

Zakhia et al. 2008), and animals fur (Suutari et al. 2010). Even the most arid habitats exposed to very high radiation levels like rocks and soil in the Atacama desert can be populated by desiccation-tolerant cyanobacteria (Tian et al. 2001; Garcia-Pichel et al. 2001; Warren-Rhodes et al. 2006).

1.2. Metabolic engineering of cyanobacteria

Today over 80% of total primary energy used globally is obtained through the combustion of coal, crude oil, and natural gas (International Energy Agency: World Energy Statistics 2015). Fossil fuels are therefore responsible for two thirds of the global greenhouse gas emissions and thus contribute significantly to global warming (Edenhofer et al. 2015). Only 10% of the energy is obtained from sustainable organic resources, mainly traditional biomass like fire wood, charcoal, agricultural residues, and animal dung (Jawahar 2015). Biofuels - renewable liquid or gaseous alternatives for fossil fuels - have only a minor share in todays energy mix, but annual production is increasing rapidly (International Energy Agency: Renewables Information 2015). Todays biofuels of the first generation are mainly produced in two ways: fermentation of sugars (bioethanol) or transesterification of oil seeds (biodiesel) (Dutta et al. 2014). Both methods require the extensive cultivation of sugarcane, corn, soybeans, rapeseed, or other oil crops, which require huge areas of arable land and high amounts of fresh water. Production of first generation biofuels is therefore limited and competes with the cultivation of animal feedstock and human food products. New approaches are therefore needed to close the global carbon cycle and curb global warming.

In recent years, cyanobacteria emerged as viable alternative to traditional methods of biomass production. Using the sun as a free and basically unlimited light source, these microorganisms can be economically cultivated either using open raceway ponds or more complex closed bioreactor systems (Jorquera et al. 2010). Compared to conventional farming, cultivation of cyanobacteria has considerable benefits. To produce a specific amount of biomass, they need less fresh water (Habib et al. 2008) or can alternatively grow in brackish, salt, or waste water (Markou and Georgakakis 2011); they can be cultivated on inarable land and are therefore in no competition to traditional food crops; and they can grow much faster and in higher density decreasing the required land area by two to three orders of magnitudes (Chisti 2007). Furthermore, cyanobacteria - especially nitrogen-fixing strains - have very low nutritional requirements. In addition to light, water, and CO_2, they only need minimal amounts of phosphate and trace metals for a sustained growth (Quintana et al. 2011; Moreno et al. 2003; Raoof et al. 2006). In consequence, cyanobacteria were investigated for the sustainable synthesis of various commodities (Lai and Lan 2015; Angermayr et al. 2015; Wang et al. 2012; Quintana et al. 2011). The list of bioproducts includes, but is not limited to:

Bioethanol Sugars synthesized by cyanobacteria can be fermented to ethanol under anaerobic dark conditions (Angermayr et al. 2009).

Biodiesel Cyanobacteria are among the most efficient producers of (un)saturated lipids and are investigated to replace oleaginous crops as source for biodiesel (Li et al. 2008).

Hydrogen Multiple cyanobacterial strains naturally produce hydrogen as a secondary metabolite, which can be used for combustion or to produce electricity in hydrogen fuel cells (Parmar et al. 2011).

Biomass Cyanobacterial biomass can be used as animal feedstock or serve as organic base material for further biotechnological processes, e.g. fermentation (Lum et al. 2013; Möllers et al. 2014; Das et al. 2015).

Bioplastic Ethylene and 3-hydroxybutyrate synthesized by cyanobacteria might replace fossil fuel based precursor for the production of various plastics (Ungerer et al. 2012; Wang et al. 2013).

Food supplements Cyanobacteria are rich in polyunsaturated fatty-acids, vitamins, as well as prebiotic, anti-inflammatory, and antioxidant substances (Raposo et al. 2013a). Specifically the most widely cultivated *Arthrospira platensis* (Spirulina) is advertised and sold as dietary supplement (Kumar et al. 2011; Ovando et al. 2016).

Other biomolecules Multiple bio-active compounds were identified in cyanobacteria and are evaluated for therapeutic, anti-fouling, and insecticide effects (Ducat et al. 2011).

Establishing economically viable synthesis of commodities in cyanobacteria most often requires optimization of growth and yield through molecular biological modification of the required pathway. Various approaches including adaptation of growth conditions, transformation of cells with alien enzyme and transporter genes, rerouting fluxes through targeted genetic knock-outs, as well as enhancing efficiency of photosynthesis and CO_2 assimilation have been discussed and applied to optimize fluxes through desired pathways (Quintana et al. 2011; Ducat et al. 2011; Wang et al. 2012; Ungerer et al. 2012; Wang et al. 2013; Raposo et al. 2013b; Angermayr et al. 2015; Savakis and Hellingwerf 2015; Lai and Lan 2015; Carmo-Silva et al. 2015; Oliver et al. 2016). To identify promising approaches, a comprehensive understanding of the cyanobacterial metabolism is crucial. Metabolic modeling, the reconstruction and subsequent evaluation of metabolic networks, has emerged as one of the most expedient strategies. In particular, flux balance analysis (FBA) proved to be a viable tool for mathematical analysis of genome-scale metabolic networks (Orth et al. 2010; Steuer et al. 2012; Cogne et al. 2011; Dal'Molin et al. 2010). Although the computational analysis is well established, reconstruction of comprehensive strain-specific metabolic networks is still a labor-intensive process, as annotations of most genomes are flawed and require manual curation. Some attempts have been made to accelerate the process of reconstruction by automated generation and completion of metabolic networks (Henry et al. 2010; Vitkin and Shlomi 2012; Caspi et al. 2014).

All cyanobacteria have photosynthetic capabilities but differ in their complexity, lifestyle, and preferred habitat. This is also reflected in their metabolism. Core metabolic processes such as glycolysis and the pentose phosphate pathway are conserved in all cyanobacterial strains, whereas the distribution of other pathways is much more fragmented. The latter includes pathways for the synthesis of compounds highly interesting for ecological and commercial application, such as ethanol

(Stal and Moezelaar 1997), hydrocarbons (Coates et al. 2014), alkane (Klähn et al. 2014), and other secondary metabolites (Dittmann et al. 2015) as well as synthesis of fatty acids (Chi et al. 2008), carotenoids (Takaichi and Mochimaru 2007), the ability to fixate atmospheric nitrogen (Zehr 2011), and many more. Therefore, to comprehend the full potential of the cyanobacterial pan-metabolisms, it is essential to understand the differences in metabolic capabilities of a wide range of cyanobacteria. However, comprehensive models for the genome-wide metabolism are still limited to a small set of selected model strains, including *Synechocystis* sp. PCC 6803 (Knoop et al. 2010, 2013; Montagud et al. 2010; Saha et al. 2012), *Synechococcus elongatus* PCC 7942 (Triana et al. 2014), *Arthrospira platensis* NIES-39 (Cogne et al. 2003; Yoshikawa et al. 2015a), and *Cyanothece* sp. ATCC 51142 (Vu et al. 2012; Saha et al. 2012). Metabolic features and differences of most other cyanobacterial strains therefore remain largely enigmatic.

A comprehensive, genome-wide understanding of metabolic processes in a variety of cyanobacteria could reveal new approaches for an efficient synthesis of valuable products. In this thesis, I addressed this knowledge gap by comparing multiple cyanobacterial strains through an exhaustive search for orthologous and thus functionally related genes. In two studies, presented in Chapters 3 and 4, I identified core enzymes common to all evaluated strains as well as shared enzymes occurring in only a subset of cyanobacteria. A thorough search for co-occurring genes revealed entire metabolic pathways and their distribution of assigned cyanobacterial strains. Subsequent simulations of automated metabolic network reconstructions demonstrated the biosynthetic capacities of individual cyanobacteria, thus supporting future research addressing metabolic engineering of cyanobacteria.

1.3. The cyanobacterial circadian clock

Due to the rotation of the earth we are exposed to a rhythmical change of the sun light's intensity. Every 24 hours we experience approximately 12 hours of light (day) and 12 hours of darkness (night). Though, the exact ratio of day and night depends on the time of the year and latitude of your geographic position. Because cyanobacteria - and all other phototrophic organisms for that matter - are dependent on sun light as their main source of energy, they are particularly affected by that rhythm. To survive phases of darkness where photosynthesis is not possible, they accumulate carbon storage compounds - mainly glycogen (Kromkamp 1987) - during the day. At night these compounds can be metabolized by means of aerobic respiration or anaerobic fermentation to produce enough energy for cell survival and various vital processes (Whitton 2012). These processes may include production of toxins, cell division, as well as DNA replication and repair (Penn et al. 2014; Holtzendorff et al. 2001). Strains of the genus *Cyanothece* use night phases to fix atmospheric nitrogen, as the required enzyme nitrogenase is susceptible to oxygen and thus incompatible with the oxygen evolving photosynthesis during the day (Sherman et al. 1998). Temporal control of biochemical processes is therefore a necessity in the life of phototrophic organisms.

Circadian clocks are endogenous biochemical mechanisms oscillating with a period

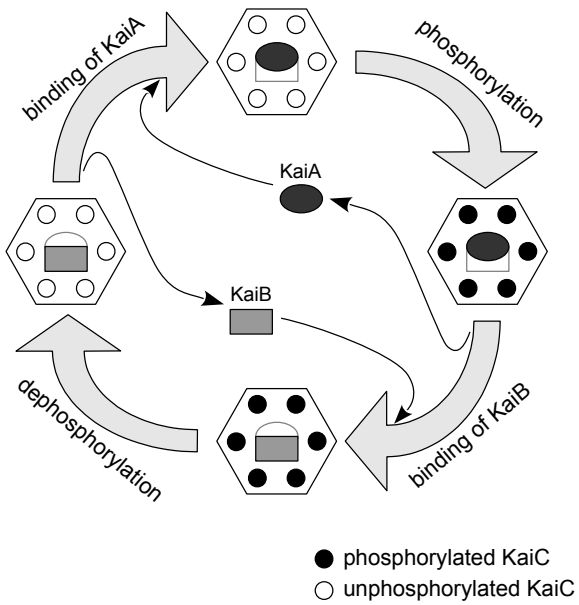


Figure 1.1.: Outline of the cyanobacterial circadian clock. KaiA binds to KaiC hexamers and triggers a multi-step autophosphorylation of KaiC. Upon full phosphorylation, KaiB binds to KaiC and inhibits the effect of KaiA, thus leading to autodephosphorylation of KaiC. Once fully dephosphorylated KaiB is released and the cycle starts again.

of roughly 24 hours. Through coupling with transcription, translation, or post-translational processes, these clocks can alter the behavior of organisms depending on the time of the day even in the absence of external time cues. Circadian rhythms have been measured in all domains of life even though the underlying mechanism is not conserved (Edgar et al. 2012). Most cyanobacteria possess a circadian clock that consists of only three proteins called KaiA, KaiB, and KaiC (Ishiura et al. 1998) - kai translates to "cycle" in Japanese. Although similar in function, the underlying mechanism is completely different from the circadian clock in humans and other animals. While the latter is based on two conjoined transcription-translation feedback loops (Ko and Takahashi 2006), cyanobacterial *kai* genes form a post-transcriptional phosphorylation/dephosphorylation-cycle at two protein sites of KaiC (Kitayama et al. 2008). Central protein of the clock is KaiC which forms a hexameric barrel-like structure (Mori et al. 2002; Hayashi et al. 2003) and has autophosphorylation as well as autodephosphorylation capabilities (Nishiwaki et al. 2000; Nishiwaki and Kondo 2012). To explain the mechanism - outlined in Figure 1.1 - in a nutshell: autophosphorylation of KaiC is stimulated by binding of KaiA (Pattanayek et al. 2008; Iwasaki et al. 2002); full phosporylation of KaiC triggers the binding of KaiB (Kitayama et al. 2003); KaiB inhibits the effect of KaiA thus prompting the dephosphorylation of KaiC; finally KaiB is released upon full dephosphorylation to complete the cycle.

The free running period of the endogenous KaiABC clock is about 24 hours. In

order to compensate for small drifts and changes in the outside world, it has to be repeatedly entrained. In cyanobacteria, three proteins play an important role in the synchronization of the circadian clock by either sensing the level of photosynthetic activity through measuring the redox state of the electron transport chain (CikA, LdpA), or directly sensing the absence of light (Pex). These proteins subsequently stimulate the phosphorylation of KaiC directly or down-regulate expression of *kaiA* leading to delayed dephosphorylation of KaiC. Four proteins take part in the read-out of the circadian clock: SasA, RpaA, LabA, and the bifunctional CikA. SasA binds to phosphorylated KaiC and transfers a phosphate group to the response regulator RpaA, which in turn activates expression of multiple genes including *kaiC*. Kinase activity of SasA is likely inhibited by LabA and CikA (Axmann et al. 2014). In addition, the topological status of the chromosome is potentially altered by the circadian clock, leading to a global oscillation of gene expression (Vijayan et al. 2009).

Overall, the cyanobacterial clock is remarkably simple yet very robust. In fact Nakajima and colleagues reconstituted the circadian oscillation in vitro by just mixing the three kai proteins in a specific right ratio and adding ATP (Nakajima et al. 2005). Despite its simplicity, this oscillatory mechanism fulfills all criteria for a full circadian clock. The oscillation is self-sustained with a period of approximately 24 hours; the oscillation is temperature compensated (Nakajima et al. 2005); and the oscillation can be entrained by changes in the environment, specifically changes in the light pattern (Rust et al. 2011).

The exact benefits of a functional circadian clock are still in debate. Though a number of hypotheses have been put forward, e.g. optimal utilization of a limited oscillating resource (light), synchronized cell-to-cell communication, or the ability to escape high UV radiations around noon (Woelfle and Johnson 2009). Disrupting the circadian clock does not lead to an obvious phenotype or changes in growth rate. Yet when growing in direct competition, strains whose internal clock is resonating with the environmental light/dark cycle outperform strains whose clocks are of different period length within a few generations (Ouyang et al. 1998; Woelfle et al. 2004). The underlying mechanism however remains enigmatic. As a matter of fact, while some cyanobacteria have multiple copies of the *kaiB* and *kaiC* (Axmann et al. 2014; Wiegard et al. 2013; Dvornyk et al. 2003), others do not possess a fully functional circadian clock. Organisms of the genus *Prochlorococcus* for example lack *kaiA*. As a result, their clock is not a self-sustained oscillator, but thought of as an hourglass-like time device that needs a daily external trigger to reset the clock (Holtzendorff et al. 2001, 2008; Axmann et al. 2009; Mullineaux and Stanewsky 2009). Even more surprising, *Gloeobacter violaceus* does not possess a single *kai* gene copy and therefore in all probability has no oscillator (Nakamura et al. 2003). Overall, a functional circadian clock seems to be beneficial, but not vital for the cell survival.

Understanding the regulatory impact of the cyanobacterial circadian clock on the level of gene expression has been a major objective during the last decade. With recent advances in transcriptomics, shrinking costs for microarrays, and the introduction of RNA-Seq (Wang et al. 2009), conducting extensive transcriptome time series became feasible. In 2005, Kucho and colleagues were the first to study circadian expression in cyanobacteria conducting a microarray time-series analysis of *Synechocystis* sp. PCC 6803 (Kucho et al. 2005), but soon were followed by multiple

other studies (Labiosa et al. 2006; Stöckel et al. 2008; Toepel et al. 2008; Zinser et al. 2009; Vijayan et al. 2009; Ito et al. 2009; Toepel et al. 2009; Straub et al. 2011; Lehmann et al. 2013; Kushige et al. 2013). Measuring gene expression levels at multiple points throughout a period of time gives an insight into the intracellular changes between day and night. Genes that oscillate with a period of approximately 24 hours in a 12 hour light/12 hour dark environment are called diurnal genes. Though they don't have to be regulated by the circadian clock because the oscillations could be induced by changes in the illumination (dusk/dawn), they indicate the severe changes cyanobacteria undergo between light and dark period. Depending on the strain, between 20 and 80 percent of the transcriptome is changing in response to the light received (Table 1.1 upper section). Genes whose rhythmic expression persists with a period of 24 hours even after the cells were transferred to a constant light setting have to be regulated by the internal circadian clock. In cyanobacteria, the number of these circadian genes is highly dependent on the investigated strain, and the applied method. It can range from less than 1.5% in *Anabaena* sp. PCC 7120 (Kushige et al. 2013) up to 64% in the circadian clock model organism *Synechococcus elongatus* PCC 7942 (Vijayan et al. 2009). Using randomly inserted luciferase assays, Liu and colleagues were able to detect almost 800 clones with detectable bioluminescence - all showing clear circadian rhythmicity (Liu et al. 1995) (Table 1.1 lower section).

As pointed out in the list of previously published studies, the absolute number of diurnal genes is fairly similar in most studied cyanobacteria, ranging from 1,133 to about 1,445, with the 2,823 rhythmic genes in *Crocosphaera watsonii* being the only exception. However, the fraction of diurnal genes can be as low as 25% in the large *Microcystis* and up to 79% in the particularly small *Prochlorococcus* strain. The opposite can be observed for the number of circadian clock-regulated genes. While in the model organism for the circadian clock *Synechococcus elongatus* PCC 7942 up to 64%, a total of 1,748 genes, oscillate in constant light, hardly any clock-regulated gene can be identified in *Anabaena* or *Synechocystis*. Even in *Cyanothece*, a cyanobacterium that undergoes massive metabolic changes during dark phases to enable fixation of atmospheric nitrogen, only about 10% of the genes are clock regulated. Furthermore, this result is particularly interesting for the case of *Synechocystis* sp. PCC 6803. This cyanobacterium not only harbors an *kaiA-kaiB-kaiC* operon closely related to that in *Synechococcus elongatus* PCC 7942, but has two additional copies of *kaiB* and *kaiC*, respectively. However, the functionality of the clock is questionable because of the low number of circadian regulated genes identified by Kucho et al. 2005.

To address the contradiction of multiple copies of clock genes leading to a neglectable number of circadian genes, I conducted a thorough study of the gene expression in *Synechocystis* sp. PCC 6803. With three different light settings - oscillating light/dark cycle, transfer to continuous light, and transfer to continuous darkness - and a sampling rate of one to two hours, this is the most comprehensive study of periodic genes not only for *Synechocystis* but for any cyanobacterial organism thus far. While I found a clear schedule of gene activation and deactivation throughout

	Strain	Light	Genes	Periodic	Reference
Diurnal genes	*Microcystis aeruginosa* PCC 7806	LD 12:12	6360	25% 1344	Straub et al. 2011
	Cyanothece ATCC 51142	LD 6:6	5359	27% 1400	Toepel et al. 2009
	Cyanothece ATCC 51142	LD 12:12	5359	30% 1445	Stöckel et al. 2008
	Synechocystis sp. PCC 6803	LD 12:12	3725	32% 1133	Lehmann et al. 2013
	Synechocystis sp. PCC 6803	LD 14:10*	3725	37% 1349	Labiosa et al. 2006
	Crocosphaera watsonii WH8501	LD 12:12	5658	47.4% 2823	Shi et al. 2010
	Prochlorococcus marinus MED4	LD 14:10	1756	79% 1403	Zinser et al. 2009
Circadian genes	*Anabaena* PCC 7120	LL	6223	1.25% 78	Kushige et al. 2013
	Synechocystis sp. PCC 6803	LL	3725	2-9% <250	Kucho et al. 2005
	Cyanothece ATCC 51142	LDLL	5359	10% ≈550	Toepel et al. 2008
	Synechococcus elongatus PCC 7942	LL	2715	29% 800	Ito et al. 2009
	Synechococcus elongatus PCC 7942	LL	2715	64% 1748	Vijayan et al. 2009
	Synechococcus elongatus PCC 7942	LL	2715	100%** 800	Liu et al. 1995

Table 1.1.: List of time series studies in cyanobacteria. Each line represents one published study giving information about strain, light regime, number of genes in genome, and fraction of oscillating genes. The light regime is abbreviated to LL for constant illumination and LD for light/dark oscillations with the given period lengths in hours. Studies are sorted according to the number of periodic genes and separated by the applied light regime. The upper part enlists time series carried out in light/dark changing settings and therefore indicates the rate of diurnal genes. The studies listed in the lower part, show the fraction of truly circadian clock controlled genes due to constant light settings. *Labiosa and colleagues used a wave-like illumination pattern, 14 hours is the total time from dawn till dusk. **The study by Liu and colleagues used a random luciferase assay approach, detecting significant expression of less than 800 different genes - all of which showed oscillating patterns. All other studies used microarray measurements to detect gene expression, thus the fraction of non-periodic genes includes all silent genes as well.

the day in oscillating light/dark conditions, I was unable to identify circadian clock controlled genes. None of the transcripts maintained rhythmic oscillations after the cells were transferred to continuous light or continuous dark conditions. Instead, I found tight regulation of the *kai* clock genes by multiple *cis*-encoded regulatory antisense RNAs. In addition, huge variations in the amount of ribosomal RNAs could be observed between light and dark periods that also dwindled in constant light conditions. To the authors knowledge, this has not been reported before.

1.4. Structure of the present thesis

In this dissertation, I will present three conducted studies addressing diurnal regulation of the global transcriptome in the cyanobacterial model organism *Synechocystis* sp. PCC 6803, as well as genetic conservation and diversity across various cyanobacterial strains. Following the general introduction, each study will be introduced in a single chapter, structured into four sections: a short specific introduction, a methods section introducing all relevant materials and techniques, a section presenting the results of all experiments, and a final discussion of the most relevant findings. The last Chapter 5 summarizes the most important results and will put the research into the context of recent publications and future developments in that field.

In the first study, presented in Chapter 2, I performed a microarray time-series analysis to investigate the expression of genes in various light conditions. With a duration of 48 hours, a time resolution of one to two hours, and three different light settings, this is the most exhaustive study on rhythmic gene expression in cyanobacteria so far. Using microarrays that covered the expression of small non-coding RNAs, I was able to shed light on the temporal regulation of various transcripts, including the *kai* clock genes. Furthermore, I observed and investigated massive diurnal oscillations in the amount of long ribosomal RNAs, which was to my knowledge not reported so far but has major implications for normalization of the microarray raw data. This study was published in Applied Environmental Microbiology (Beck et al. 2014).

The second study is discussed in Chapter 3 and addresses similarities and differences in the genomic information of multiple cyanobacterial strains, while specifically focusing on variations in their metabolic networks. With an exhaustive reciprocal BLAST search, I identified and clustered orthologous genes in sixteen selected cyanobacteria. I was able to make reasonable estimates for the sizes of the core genome - genes that are common in all cyanobacteria - as well as the pan-genome, comprising all genes. Furthermore, I identified and discussed conservation and diversity in various core metabolic pathways of cyanobacteria. This study has been published in BMC Genomics (Beck et al. 2012).

In the third study, presented in Chapter 4, I expanded on the idea of the reciprocal BLAST method introduced in the previous chapter. Now including 77 cyanobacterial strains allowed me to refine the number of core, shared, and unique genes. I developed and applied a method to identify and group genes co-occurring in similar subsets of genomes, which could not be explained by lateral gene transfer or pure chance alone. The joint conservation of two or more genes implies a common function. I identified

various groups of co-occurring proteins that are involved in a joint metabolic pathway, co-participate in the same intracellular structures, or collectively form one protein complex. Many of the co-occurring genes are not located in close proximity on the cyanobacterial genomes and therefore can not be identified by just analyzing the genome structures. With the SimilarityViewer I developed an easy-to-use graphical computer tool for the identification of genes co-occurring in the studied strains. By combining the enzyme annotation of multiple resources, I was able to automatically reconstruct the metabolic network of all studied organisms. Subsequent simulation of the metabolisms revealed the similarity and diversity in the biosynthetic capabilities of the strains, which were in good accordance with published biochemical studies. Parts of this study have been submitted and are available as a preprint (Beck et al. 2018).

Remark on word usage

I note that from now on, in accordance with the standard scientific protocol, the personal pronoun "we" will be used to indicate the reader and the writer, or the author and his scientific collaborators.

Chapter 2.

Diurnal expression of protein-encoding genes and non-coding RNAs in *Synechocystis* sp. PCC 6803

2.1. Introduction

Cyanobacteria in general use sunlight as their main source of energy and to synthesize all essential organic components required for continuous growth. Due to the rotation of the earth, however, they are subject to daily fluctuation of the perceived light. To cope with the alternation of day and night, most cyanobacteria incorporate a robust timing mechanism. This circadian clock is composed of three proteins KaiA, KaiB, and KaiC, whose interaction results in an oscillation of phosphorylation and dephosphorylation of KaiC with a period of approximately 24 hours as outlined in Figure 1.1 on page 16 (Nakajima et al. 2005). This mechanism was first identified and is well studied in *Synechococcus elongatus* PCC 7942, subsequently shortened to *Synechococcus* (Ishiura et al. 1998; Kitayama et al. 2008). However, not all cyanobacteria possess all three *kai* genes, resulting in highly damped or hourglass-like oscillators (Johnson et al. 2017). *Synechocystis* sp. PCC 6803 (hereafter *Synechocystis*) is a model organism for the study of cyanobacteria and more generally oxygenic photosynthesis. It was the first phototrophic organism with a fully sequenced genome (Kaneko et al. 1996). Interestingly, the genome of *Synechocystis* contains multiple copies of the three circadian clock genes. In addition to a *kaiA-kaiB-kaiC* operon most similar to that in *Synechococcus*, another *kaiB-kaiC* cluster (*kaiB2/kaiC2*) and two separated *kaiB* and *kaiC* genes (*kaiB3/kaiC3*) were identified. All three copies of the central *kaiC* seem to be structurally intact, as we identified phosphorylation activities of all proteins. However, while KaiC1 only phosphorylated in the presence of KaiA, as was reported for KaiC of *Synechococcus*, KaiC2 and KaiC3 showed autophosphorylation activity (Wiegard et al. 2013). The latter can also be observed for the hourglass-like timing mechanism in *Prochlorococcus* sp. strain MED4, which possesses homologues of KaiB and KaiC, but no KaiA (Axmann et al. 2009). Cellular function related to oscillatory gene expression has been described for *kaiC* homologues in non-cyanobacterial organisms (Ma et al. 2016; Loza-Correa et al. 2014), yet in a recently published study, we could not identify a single organism outside the cyanobacterial phylum that possessed a homolog of *kaiA* (Schmelling et al. 2017). The exact function of the additional Kai proteins in *Synechocystis* and other cyanobacteria is currently unknown. Possible explanations include fine-tuning

or improved robustness of the circadian clock (Aoki and Onai 2009).

Despite multiple copies of clock genes, however, circadian expression in *Synechocystis* was reported for only a small fraction of the genes. In 2005, Kucho and his colleagues selected *Synechocystis* to perform the first microarray time-series study on any phototrophic organism. Depending on the stringency of their detection method, only 2% (54 genes) or 9% (237) of the genes showed circadian oscillations in continuous light conditions (Kucho et al. 2005). In contrast, microarray time-series experiments with *Synechococcus* identified between 29% and 64% of the genes regulated by the circadian clock (Ito et al. 2009; Vijayan et al. 2009).

To address this discrepancy, we expanded the previous study by Kucho and colleagues and performed a refined microarray time-series experiment. With a duration of 48 hours, a sampling rate of two hours, and three different illumination conditions (depicted in Figure 2.1) including a persistent 12-h-light-12-h-dark rhythm, transfer to continuous light, and transfer to continuous dark, this is the most thorough time-series analysis in any cyanobacterial organisms so far. In our study, we used custom-made microarray chips covering all transcripts identified in a previous transcriptomics study (Mitschke et al. 2011), including non-coding small RNAs. We specifically looked at the regulation of *kai* genes through asRNAs, small non-coding RNA transcripts encoded on the opposite DNA strand of the regulated gene. AsRNAs are involved in multiple intracellular processes of bacteria (Georg and Hess 2011) and can amplify the degradation of transcripts by forming a double-stranded RNA with their target RNA that is degraded by specific RNases. Alternatively, asRNA can bind at the 5' or the 3' end of a transcript and thereby increase termination efficiency or protect the target RNA from decay, respectively (Georg et al. 2009). In our time-series experiments we identified both positive and negative regulation of the *kai* genes.

When quantifying the purified RNA, we observed strong diurnal oscillations of the total amount of RNAs in light-dark conditions. This challenges the common assumption that the quantity of RNA is related to the cell activity and growth rate (Kjeldgaard and Kurland 1963; Poulsen et al. 1993; Binder and Liu 1998). Interestingly, fluctuations were not equally distributed among all types of RNA but rather specific for long ribosomal RNAs. However, because of the large share of ribosomal RNAs (Bremer and Dennis 2008), this significantly affects the amount of total RNA as well. In consequence, fluctuating amounts of RNA also have huge implications for the chip-to-chip normalization of microarray gene expression data, as most common normalization methods assume some sort of invariant mRNA levels. Yet, because of variation of highly abundant non-coding RNAs, this does not hold true for our data set. Furthermore, the amounts of ribosomal RNAs could not be measured with the microarrays due to saturation effects. Normalization under these circumstances was addressed in a previous study of *Synechocystis* by Lehmann et al. 2013. To the authors' knowledge, such strong oscillations of ribosomal RNA have not been reported thus far, and possible implications for the cyanobacterial metabolism will be discussed in this chapter.

2.2. Materials and methods

2.2.1. Growth conditions

The motile and glucose-tolerant wild type of *Synechocystis* sp. PCC 6803 obtained from Sergey V. Shestakov of Moscow State University in Russia, was cultivated photoautotrophically in BG-11 medium at 30 °C under continuous illumination with white light at 80 $\frac{\mu\text{mol photons}}{\text{m}^2\ \text{s}}$ and a constant stream of air. Cell density was determined by measuring the optical density at a wavelength of 750nm (OD_{750}) and manual counting. The culture was kept in log growth phase and below an OD_{750} of 1.0 by regular dilution. Three days prior to the time-series experiments, the cultures were diluted to a specific volume with OD_{750}=0.4 and transferred to a 12-h-light-12-h-dark cycle for synchronization. The time point of the third transition from dark to light was denoted as time zero and all times were related to this time point. To maximize the number of light/dark transitions, sampling of times series started at 5.5 hours (h). Fitness of the cultures was monitored during the time-series experiments by measuring growth rate and concentrations of chlorophyll and phycocyanin. The recorded physiological data indicated that the cells grew under unstressed conditions.

2.2.2. Sampling and RNA extraction

The batch cultures were cultivated under three different light regimes, sketched in Figure 2.1: permanent 12-h-light-12-h-dark cycle (LDLDL), transfer to continuous light (LDLLL), or transfer to continuous dark (LDDDD). Starting at time point 5.5h, 15 ml of culture were sampled for 48 hours every 2 hours with one additional sample 30 minutes after each transfer or subjective transfer from dark to light. In time-series experiments regarding the concentration of ribosomal RNA (rRNA), further samples were taken every 10 minutes in the time span between 30 minutes before and 30 minutes after transition of (subjective) dark to light, plus an additional sample 45 minutes after dark-light transition. With this setup, we ensure a high sampling rate throughout the 48 h time-series experiment with a particularly dense sampling at the most critical phase at the beginning of the day.

Each sample of cyanobacterial culture was filtered through Supor 0.45 μm membrane filters (Pall), immediately mixed with TRIzol™ reagent (Thermo Fisher Scientific), and frozen in liquid nitrogen. Samples were stored at -20 °C for no more than 2 days. To purify the RNA, frozen samples were directly heated to 65 °C for 5 min, mixed with 0.2 ml chloroform per ml of TRIzol™, and incubated for another 15 min. Cell lysis was supported by vortexing and phases were separated by centrifugation at maximum speed for 10 minutes at 4 °C. RNA in the supernatant was precipitated by adding one volume of isopropanol and treated with RNase-free TURBO™ DNase (Ambion®) by following the manufacturer's protocol, which resulted in 40 μl purified RNA solution.

The total amount of RNA was measured with a NanoDrop™ ND-1000 spectrophotometer (Thermo Scientific™), normalized by the OD_{750}, and divided by 15, resulting in the concentration of RNA in 1 ml culture with OD_{750}=1.

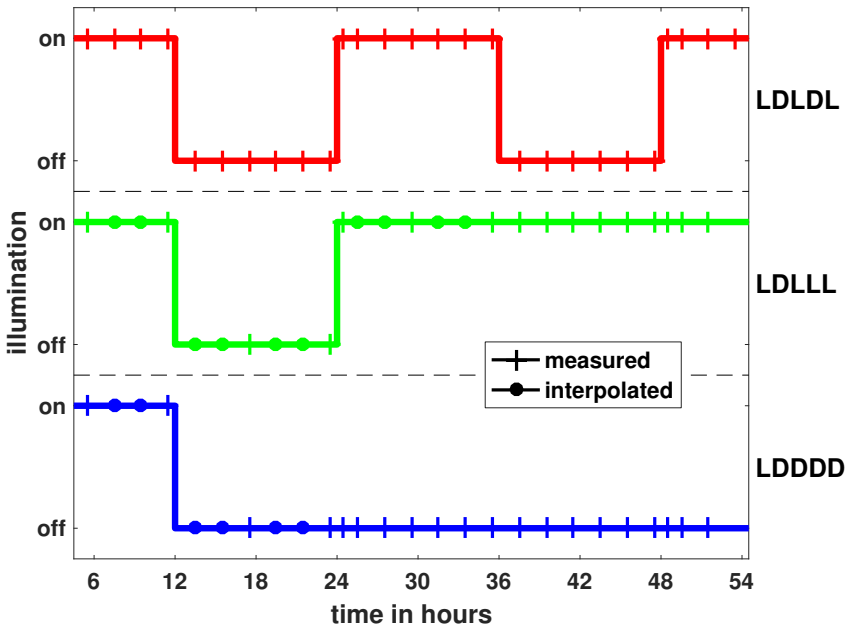


Figure 2.1.: Illumination scheme and sampled time points of the three different time-series experiments. We performed three time-series experiments with three different light settings: oscillating light/dark periods (LDLDL), transition to continuous light (LDLLL), and transition to continuous dark (LDDDD). In the LDLDL series, samples were taken every 2 hours starting at 5.5 hours after the light onset of the first day, with an additional sample 30 minutes after each subsequent switch from dark to light (hours 24.5 and 48.5). In both other experiments, time points of the overlapping periods were partially interpolated using the data of the LDLDL experiment (marked with dots).

2.2.3. Quantification of ribosomal RNA

To analyze the composition of RNA and especially the concentration of rRNA, purified RNA was analyzed using the microfluidic electrophoresis system Agilent 2100 Bioanalyzer. 1 μl of purified RNA solution was prepared with the RNA 6000 Nano Kit according to the manufacturer's instructions (Panaro et al. 2000). In short, RNA was labeled with a fluorescent tag and complemented with a specific RNA marker. In addition, the ladder, an artificial solution of RNA molecules with specific length and quantity was labeled. Twelve samples and the ladder were pipetted onto the gel of one RNA Nano chip and quantified in the Bioanalyzer. Unaligned raw data provided by the Agilent 2100 Expert Software was exported to automate the analysis of rRNA concentrations. One electropherogram is exemplarily shown and described in Figure 2.2. The raw data was normalized and rRNA concentrations were determined as follows. First, the raw fluorescence data was divided by the migration time to compensate for higher fluorescence intensities of larger RNAs caused by slower passing of the detector. Second, the migration time of each sample was normalized

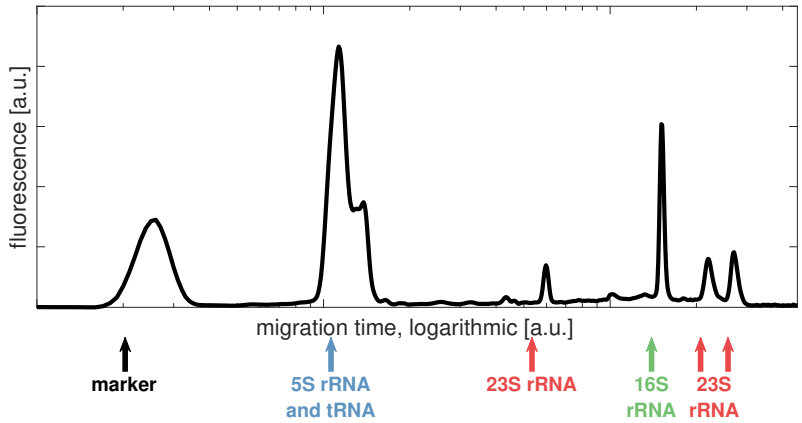


Figure 2.2.: Raw data for one sample generated by the Agilent 2100 Bioanalyzer. The x-axis shows the migration time of labeled RNAs through a gel, which is linearly related to the fragment size. Values on the y-axis are fluorescence intensities of the RNA fragments when passing the detector and a direct function of their quantity. Due to their abundance, peaks of ribosomal and transfer RNAs are clearly identifiable. The first peak belongs to the marker, an artificial RNA added according to the protocol for technical reasons. Transfer RNA (70-94 nt) and 5S ribosomal RNA (119 nt) can not be separated due to their similar sizes (blue arrow). Cleavage of 23S ribosomal RNA after maturation (Doolittle 1973) results in two fragments of around 500 and 2,400 nt that can be identified in addition to the 2,883 nt uncleaved rRNA. Contrary to Doolittle, we observed no diurnal or circadian oscillation of the fraction of cleaved 23S rRNA (Doolittle 1973). Concentrations of 23S rRNA were calculated by summing the amount of all three fragments. 16S rRNA has a size of 1,489 nt. Comparing the data with the electropherogram of a standardized RNA ladder, we were able to calculate exact sizes and concentrations of the RNA species.

by the migration time of the marker, which has a defined length of 25 nucleotides. Third, rRNAs were detected by identifying the six largest peaks corresponding to the marker, 5S rRNA and tRNAs, 16S rRNA, and the cleaved and uncleaved fragments of 23S rRNA. The background RNA concentration at each peak, corresponding to non-ribosomal RNA with the same length, was estimated by linear regression between left and right boundaries of the peak and subtracted from the area under the peak. Fourth, areas under the fluorescence curve were translated to RNA concentrations by also calculating the area under the peaks for the ladder on the same chip. The ladder aliquots have a defined concentration of 150 $\frac{\text{ng}}{\mu\text{l}}$ and therefore allow to determine the rRNA concentrations. Finally, we summarized the whole area under the curve and calculated the amount of total RNA for all samples as well.

For each electropherogram, we visually checked the correct detection of ribosomal peaks. Samples with no or clearly degraded RNA were eliminated from further analysis. The calculated concentrations of rRNA were in agreement with the concentrations exemplarily determined with the Agilent 2100 Expert Software. The concentrations of total RNA were also consistent with the values determined with the Expert Software and the NanoDrop™ spectrophotometer.

2.2.4. Microarray design and hybridization

Gene expression of *Synechocystis* was analyzed using custom-made 44K Agilent RNA microarrays. Each chip contains 20,443 probes for 3,352 protein-encoding genes, 1,931 asRNAs, and 620 other non-coding RNAs of the *Synechocystis* chromosome (NC_000911.1), as well as the plasmid pSYSA (NC_005230), identified by Mitschke et al. (Mitschke et al. 2011). All probes were spotted two times on each chip serving as technical replicates. Full information on the microarray design is stored in the Gene Expression Omnibus (GEO) database [www.ncbi.nlm.nih.gov/geo/] under the accession number GPL15867.

Purified RNA of all three light regimes LDLDL, LDLLL, and LDDDD were hybridized to microarrays. Two micrograms of RNA were directly labeled with the Cy5 dye by using Kreatech's ULS™ labeling kit for Agilent gene expression arrays. A sample of 1.5 μg labeled RNA was hybridized to each microarray following the Agilent protocol for single-color microarrays and in accordance with Georg et al. 2009. Biological replicates from time points 11.5 and 23.5 of the LDLDL time series were hybridized on two separate microarrays and confirm a good reproducibility of the results. The microarrays were digitalized with an Agilent G2505B microarray scanner, using Agilent's Feature Extraction software 10.7.3.1 and protocol GE1_107_Sep09. The raw data of each spot was extracted with the R-package *limma* (Smyth 2005) and calculated as the medium intensity of the spot minus the median of the corresponding background intensity.

RNA from ten time points of the LDLLL time series overlapping with the LDLDL series (7.5, 9.5, 13.5, 15.5, 19.5, 21.5, 25.5, 37.5, 31.5, and 33.5 h) as well as four time points of the LDDDD series (7.5, 9.5, 13.5, and 15.5 h) were not hybridized (see Figure 2.1). These missing data points were instead interpolated with the corresponding time points from the LDLDL time series using the raw chip data. For that, two intensity values for each spot of the LDLDL time series were adjusted to the according measured time point of the LDLLL series. Missing data between these two points in the latter series were than completed with the appropriate adjusted data from the former series. The same was done for unmeasured time points of the LDDDD series.

The raw microarray data was deposited in the GEO database under accession number GSE47482.

2.2.5. Data normalization and clustering

Synechocystis exhibits strong oscillations in the total amount and composition of RNA, which will be discussed in Section 2.3.3. This adds strong bias when standard normalization methods, such as quantile normalization and median polishing, are applied to the data, as these assume somewhat similar quantities in the majority of RNAs. Moreover, the intensities of rRNAs could not be detected reliably because of saturation effects. These factors were considered in a previous study comparing several normalization approaches and data analysis methods (Lehmann et al. 2013). As a result of the study, we applied the recommended method and based the normalization on a least oscillating set (LOS) of probes in the LDLDL time series. Discrete Fourier transformation (talk) was applied to the time course of each probe

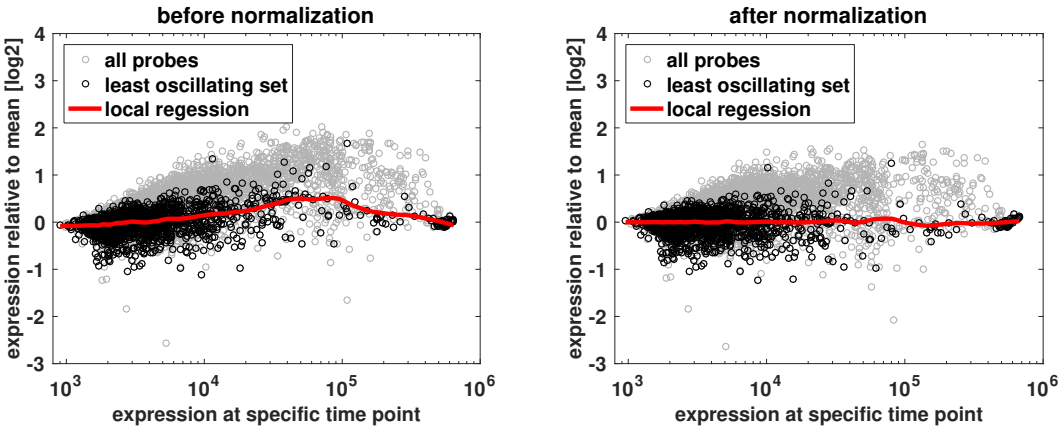


Figure 2.3.: Normalization of microarray data with least oscillating probe set. Depicted is the expression of all probes at the exemplary time point 5.5 h relative to the mean expression throughout the LDLDL time series before (left) and after (right) the normalization. LOS probes are shown in black circles and the LOWESS regression of the LOS expression is indicated by a red line. All values are normalized by the regression curve, thus minimizing the relative expression of LOS probes.

and 500,000 random permutations. The combined power of the 24-h oscillations and their two first harmonics, oscillations with periods of 12 and 8 hours, was then calculated and an estimated likelihood (e-value) was computed as the fraction of permutations with oscillatory power above the actual measured data. Probes with an e-value greater than 0.7 (representing about 17% of all probes) were selected as the LOS. For each time point, a local regression curve (locally weighted scatterplot smoothing (LOWESS) with a smoothing parameter of 0.2 (Cleveland and Devlin 1988)) was fitted to the raw data of all LOS probes and their average expression value over the whole LDLDL time-series experiment, as depicted in Figure 2.3. All expression values at the specific time point were normalized regarding these curves. For reasons of comparability, data of the LDLLL and LDDDD time series were also normalized to the average value of the LDLDL series. Finally, the data of all probes corresponding to the same gene were averaged to a single gene-specific expression value. The oscillation of each transcript was once again quantified by calculating the DFT and comparing the combined power of 24, 12, and 8 hour oscillations with the power of 500,000 random permutations. Transcripts with an e-value of 0.01 or less were considered as significantly oscillatory. The phase of the 24 hour oscillation was used to calculate the peak time of gene expression.

Upon the recommendation of Lehmann et al. 2013, we selected the *flowClust* algorithm (Lo et al. 2009) to group all protein-encoding genes based on their expressions in the LDLDL time series. This model-based clustering method is available as a component of the Bioconductor package for R [www.bioconductor.org] and was applied with standard parameters. Visual inspections of clusterings with 2 to 15 groups revealed a most promising separation when the profiles were divided into 10 groups.

2.2.6. Gene ontology enrichment analysis

All gene ontology (GO) annotations of *Synechocystis* were gathered from the gene ontology database [www.geneontology.org] (Harris et al. 2004). Enrichment of GO terms within groups of genes was calculated using Fisher's exact test.

2.3. Results

Due to the perpetual alternation of day and night, cyanobacteria have to adapt to continuous fluctuation of their main source of energy, light. Most cyanobacterial strains rely on an endogenous biological clock composed of three proteins KaiA, KaiB, and KaiC to tell the time of the day. Possessing a circadian clock that matches the environmental cycle leads to a significant improvement of reproductive fitness (Ouyang et al. 1998). However, the regulatory impact of the clock in *Synechocystis* remains vague as Kucho and colleagues reported that only 2 to 9 % of the gene transcripts showed circadian rhythm under constant light conditions (Kucho et al. 2005). To address that uncertainty we conducted three time-series studies using custom designed high density microarray chips covering the expression not only of 3,352 protein-coding genes but of 2,251 non-coding transcripts as well. The time series comprise three different light regimes, perpetual oscillation of light and dark (LDLDL), shift to continuous light (LDLLL), and shift to continuous darkness (LDDDD). Samples were taken every two hours with an additional sample 30 minutes after the (subjective) dawn and cells were closely monitored throughout the experiments to ensure unstressed growth. With this set-up, we aimed at a thorough understanding of diurnal (responding to light changes) and circadian (controlled by an endogenous clock) rhythms.

2.3.1. Diurnal gene expression

Rhythmic expression of each transcript was identified by comparing the 24, 12, and 8 hour oscillations within each time-series experiment with 500,000 random permutations. Transcripts with an e-value below 0.01 were classified as significantly oscillating. Using our threshold we identified 27% of all genes oscillating under light-dark conditions in the LDLDL series. However, oscillations disappeared almost entirely upon transfer to continuous light or dark conditions. Only 113 or about 2% of the transcripts showed significant oscillation during the LDLLL time series and only 54 (1%) transcripts during the LDDDD experiment. None of the transcripts showed consistent oscillations under all three light regimes. We therefore conclude that no transcript is actually controlled by an endogenous circadian clock as the 24 hour rhythm would have to persist in all light conditions. Rhythmicity during the LDLDL experiment rather points to a diurnal response of the expression to changes in the perceived light.

Because *Synechocystis* possesses multiple copies of *kai* genes, now the question arises of the regulatory impact of clock-related genes in daily gene expression. To test this, all protein coding genes were clustered into 10 groups, based on their expression profiles in the LDLDL time series using flowClust (Lo et al. 2009). The

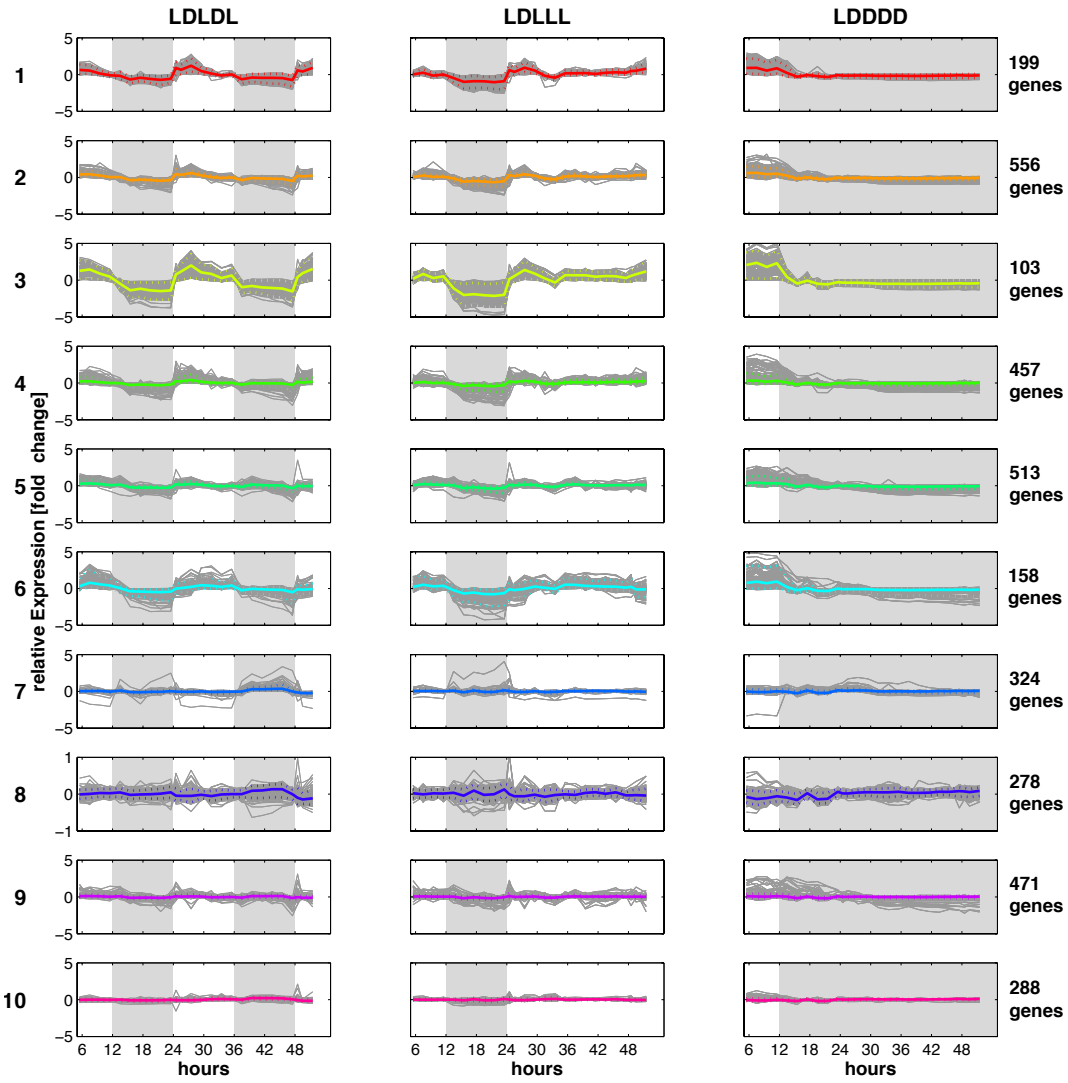


Figure 2.4.: Clustered expression profiles of protein-coding genes under three different light regimes. All protein-coding genes were clustered into ten groups according to their expression profile in LDLDL. Clusters are sorted by the oscillatory phase times, starting with cluster 1 peaking right after the onset of light, to cluster 8 peaking late during the night. Clusters 9 and 10 consist of completely arhythmic transcripts. For each gene, the relative logarithmic expression throughout the 48 hour experiments is depicted in gray lines. Solid colored lines show the average transcription within a cluster, while dashed lines indicate the 5% and 95% quantiles. Light gray backgrounds indicate dark phases during the experiments. The number of genes in each cluster is given on the right. Reprinted from Beck et al. 2014.

clusters were sorted by their mean peak time in LDLDL and are depicted with the expressions throughout all three experiments in Figure 2.4. The first six clusters consist of genes peaking during the day, right at the onset of light (cluster 1) up until the end of the day (cluster 6). Clusters 7 and 8 are made up of genes upregulated during the night, while arhythmic genes in clusters 9 and 10 had no clear peak time. As shown by the diagrams, rhythmic expression vanishes in constant light conditions. Even more interesting, expression in constant conditions matches the expression in the respective phases during the LDLDL time series. Genes upregulated during the day (cluster 1-6) are permanently upregulated in constant light (LDLLL) but downregulated throughout constant darkness (LDDDD). Again, this finding indicates strong influence of the perceived light on the expression of genes, rather than a circadian clock. This is also underlined by the fact that, while genes are strongly upregulated immediately (less than 30 minutes) after the onset of light, no genes are upregulated in anticipation (>30 minutes) of the shift to light. Cells seem to be unable to predict and prepare for the beginning of the day.

Nonetheless, we observed a distinct schedule in the upregulation of functionally related genes. For each cluster we determined the GO annotation of associated genes and calculated the enrichment of GO terms using Fisher's exact test. The most significant enrichments are listed in Figure 2.5. Cluster 1, peaking immediately after the shift to light, is highly enriched with genes related to energy production (ATP synthase) and the synthesis of new proteins (translation, ribosome synthesis). These processes are followed by the activation of genes necessary for amino acid biosynthesis (cluster 2, branched chain family amino acid metabolism) and the upregulation of the entire photosynthesis apparatus including photosystems I and II, phycobilisome light harvesting antennae, and the electron carrier complex cytochrome b_6f (cluster 3). The latter cluster also includes a significant number of genes related to the fixation of atmospheric carbon dioxide. Cluster 4 contains multiple genes associated with modification of proteins but also all genes required for the synthesis of molybdopterin, a molybdenum binding co-factor essential for a number of enzymes including nitrate reductase, and therefore vital for the cellular assimilation of nitrate (Rubio et al. 1999). Interestingly, cluster 5, peaking later in the day, and cluster 2 are significantly enriched with genes associated with the obscure regulation of transposons and hypothetical genes. This indicates a large group of genes that are upregulated during the day and are therefore likely involved in phototrophic growth, but of which we have so far only very limited understanding of their biological functions.

Genes peaking in the second half of the day are assembled in cluster 6 and are often associated with processes essential for the generation of energy in the absence of light. This includes enzymes involved in the pentose phosphate pathway, responsible for the catabolism of various carbohydrates, such as glucose-6-phosphate dehydrogenase (GPD) and 6-phosphogluconate dehydrogenase (6PGD), as well as also genes involved in respiration, like NADH dehydrogenase and cytochrome c oxidase. Clusters 7 and 8 consist of genes upregulated during the night and are significantly enriched with genes associated with transmembrane transport as well as DNA replication and repair. Genes comprising the final two clusters 9 and 10 show no clear rhythmicity but have constant expression levels. They are often involved in central metabolic

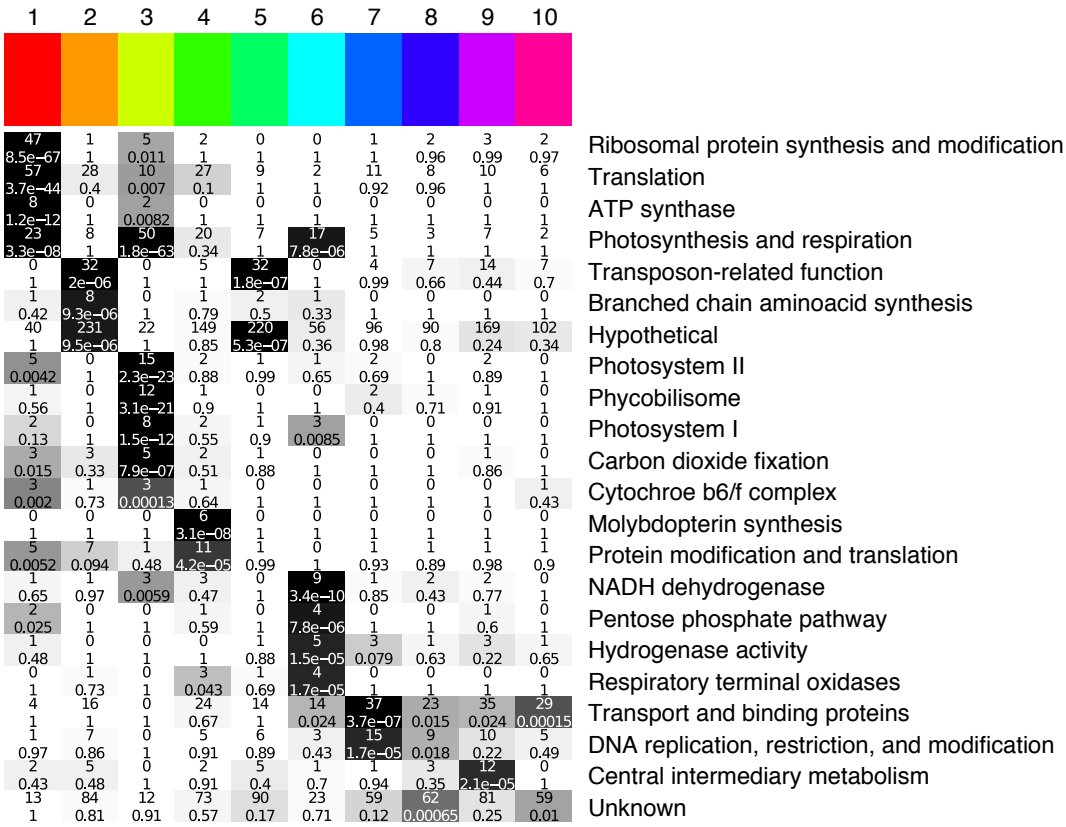


Figure 2.5.: Diurnal schedule for the expression of functionally related genes. Rows of the table correspond to specific biological functions (GO terms) indicated on the right, while columns correspond to the ten cluster of co-regulated genes shown in Figure 2.4. Clusters are sorted by their peak times starting at the onset of light (cluster 1) to late in the day (cluster 6). Genes upregulated during the night are assembled in clusters 7 and 8, while clusters 9 and 10 consist of arrhythmically expressed genes. Each cell shows the cluster-specific number of genes associated with the respective biological function, as well as the enrichment calculated with the one-tailed Fisher's exact test. P-values were adjusted using the Bonferroni correction implemented in the MATLAB *multcompare* function. For ease of presentation, the background of each cell is shaded according to the significance of the enrichment. Highly significant terms (p-value$<$0.001) appear in black or dark grey color with white text, while insignificant enrichments have white or light gray background and black text. Adapted from Beck et al. 2014.

processes and transport.

Overall, our enrichment analysis revealed clear co-regulation of functionally related genes. In addition, we identified a tight schedule for the activation of specific processes that is consistent with the phototrophic lifestyle of *Synechocystis*. Genes upregulated early in the day are most likely involved in translation, photosynthesis, carbon fixation, and amino acid biosynthesis, thus ensuring maximal phototrophic growth. Towards the end of the light period, genes involved in respiration and catabolism are activated, to prepare the cells for the upcoming night. Most processes involving DNA repair and replication as well as transport are shifted to the dark period to avoid high oxidative stress and transmembrane gradients caused by photosynthesis. In contrast to the observation of a clear diurnal schedule, the cell cycle likely is not synchronized to the light-dark oscillations. Genes involved in cell division were found in various clusters throughout the day including cluster 1 (minD), cluster 4 (minC and ftsZ), cluster 5 (zipN), and cluster 10 (minE) (Mazouni et al. 2004). We finally conclude that *Synechocystis* likely possesses an hourglass-like timing mechanism to temporally organize gene expression. However, the oscillation is not self-sustaining and a daily external cue (likely the transition from dark to light) is needed to reset the clock.

2.3.2. Expression and regulation of clock-related genes

Although *Synechocystis* lacks persistent oscillations, its genome contains multiple copies of *kai* clock genes. In addition to the standard kaiABC cluster found in various cyanobacteria such as *Synechococcus*, one kaiBC cluster, one single *kaiB* and one single *kaiC* can be found. Looking at the genomic data as depicted in Figure 2.6, it is apparent that all *kai* genes, except *kaiC2* and *kaiB3*, are regulated by at least one cis-transcoded asRNA. Nonetheless, we identified identical low-amplitude diurnal expression for *kaiA*, *kaiB1*, and *kaiC1*, indicating that these genes form an operon. Transcription of all three genes soars right at the onset of light and decreases in the second half of the light phase. The same is true for *kaiB2* and *kaiC3*, however, we could not find rhythmic expression for *kaiC2* and *kaiB3*. Although the open reading frames of *kaiB2* and *kaiC2* are separated by less than 20 nt, different expression profiles imply separate promotors for both genes rather than a combined operon structure. Rhythmic expression of the *kai* genes ceased when the cells were transferred to continuous light or darkness.

AsRNAs are small non-coding RNA transcripts encoded on the opposite DNA strand of the regulated gene. They are involved in multiple intracellular processes of bacteria (Georg and Hess 2011) and can amplify the degradation of transcripts by forming a double-stranded RNA with their target RNA that is degraded by specific RNases. Alternatively, asRNA can bind at the 5' or the 3' end of a transcript and thereby increase termination efficiency or protect the RNA from decay, respectively (Georg et al. 2009). We found evidence for both mechanisms in our data, as shown in Figure 2.6. *KaiA-as1* and *kaiBC-as1* binding at the 3' ends of *kaiA* and *kaiC1*, respectively are positively correlated with the transcripts of the kaiAB1C1 cluster (*kaiA* vs. *kaiA-as1*: Spearman's ρ=0.55, p-value of permutation test: $<$0.01) and therefore likely stabilize transcription of the clock genes. In contrast, asRNAs bind-

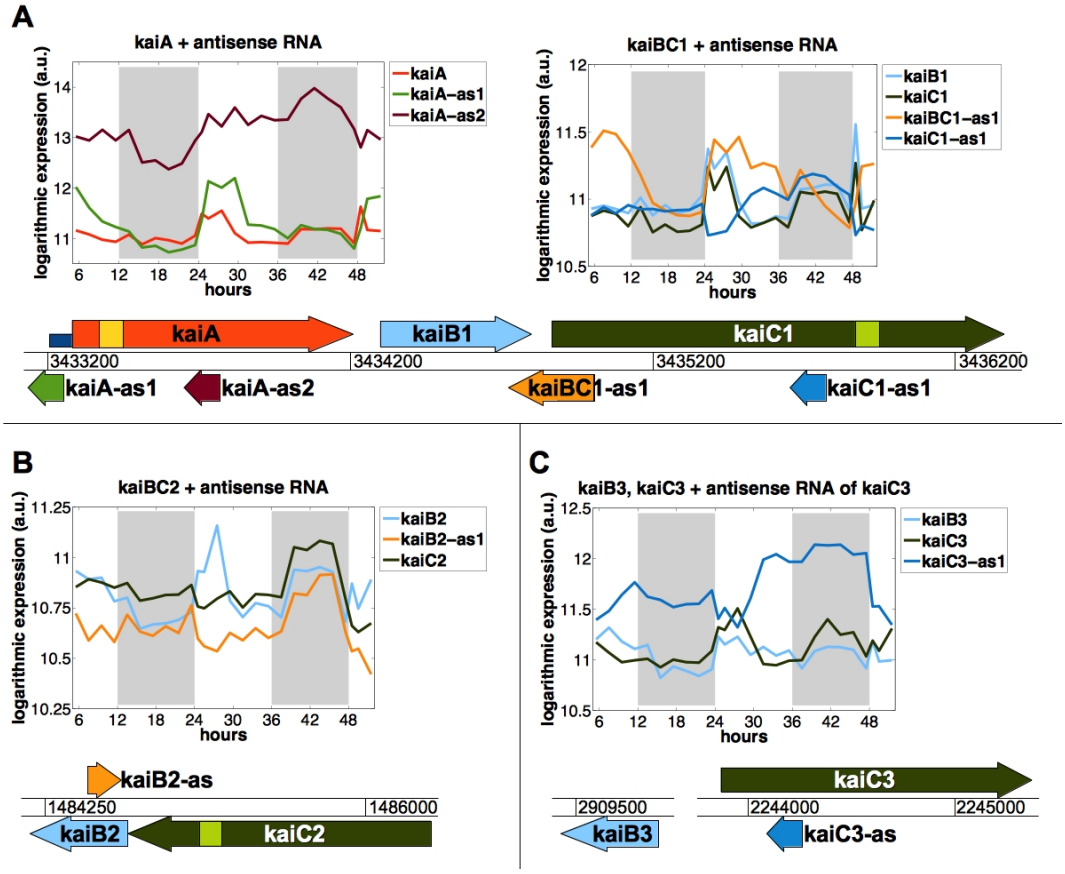


Figure 2.6.: Genomic organization and expression of homologous *kai* clock genes and their regulatory asRNAs in light-dark conditions. This figure depicts the genome positions of all core clock genes, namely the *kaiA-kaiB1-kaiC1* cluster (**A**), the *kaiB2-kaiC2* cluster (**B**), as well as the disjoined *kaiB3* and *kaiC3* (**C**), and their respective asRNAs. ORFs and reading direction are shown by colored arrows. Numbers on the strand between sense and antisense transcripts indicate the position on the forward DNA strand. Light boxes within *kaiA*, *kaiC1* and *kaiC2* point to alternative transcription start sites as identified by Mitschke et al. 2011. The small blue box at the 5'-end of *kaiA* indicates the regulatory 5'-UTR opposed by the *kaiA-as1*. Graphs above the sketched genome sections show the expression of clock genes and regulatory asRNAs throughout the LDLDL time series in matching colors. Gray boxes in the background indicate dark phases. *KaiA*, *kaiB1*, and *kaiC1* have similar expressions and are likely collectively upregulated by *kaiA-as1* binding at the promotor upstream of *kaiA*. In contrast, *kaiB2* and *kaiC2*, although being in close proximity on the genome, have separate promotors resulting in different expression patterns. *KaiB3* and *kaiC3* can not be found in close proximity on the genome and show dissimilar expression. In addition, *kaiC3* seems to be highly repressed by the binding of *kaiC3-as*. Reprinted from Beck et al. 2014.

ing in the center of ORFs such as *kaiC1-as1*, *kaiB2-as*, and *kaiC3-as* are negatively correlated (*kaiC3* vs. *kaiC3-as*: ρ=-0.68, p-value<0.01) and thus promote degradation of the *kai* transcripts. *KaiA-as2* binding in the center of *kaiA* shows neither positive nor negative correlation but has, in contrast to all other clock-related transcripts, a particularly high expression. All *kai* genes in general have a low expression rate compared to the average expression of all genes and highest observed amplitudes are in the range of a 0.5 fold increase. In contrast, amplitudes reported for *kai* genes in *Synechococcus* are in the range of two fold changes (Vijayan et al. 2009). However, absolute expression values in microarrays have to be treated with the utmost caution.

The extensive regulation of the circadian clock genes, especially the *kaiA-kaiB1-kaiC1* cluster, by asRNAs is surprising. In particular, as the regulation results in a low amplitude expression of the genes. We therefore argue that *Synechocystis* might have evolved a facultative clock system that can be switched via asRNAs regulation between an hourglass-like timing mechanism and a fully functional self-sustaining circadian oscillator, depending on environment and growth conditions. Conditionality of the circadian clock was also observed in *Synechococcus* at low temperatures, where the clock is deactivated to increase overall fitness (Xu et al. 2013a).

2.3.3. Oscillation of total and ribosomal RNA

When conducting the LDLDL time-series experiment, we measured the RNA concentration and observed a strong diurnal oscillation of total RNA per volume of culture, as shown in Figure 2.7A. During the night, the total amount of RNA in one milliliter culture with an OD_{750} of one raises from around 200 ng to 350-400 ng, thus almost doubling the concentration. With the onset of light, however, the amount rapidly drops back to the 200 ng measured at the beginning of the night. This finding was very surprising as the concentration of RNA was previously assumed to be a function of growth rate (Kjeldgaard and Kurland 1963; Poulsen et al. 1993), while our cells show almost no growth during the night. Modest accumulation of RNA during the night could be observed for *Synechococcus* sp. PCC 6301 but was attributed to delay of cell devision during the night (Lepp and Schmidt 1998).

To find out whether all RNA species oscillate equally, we used a microfluidic electrophoresis system (Agilent 2100 Bioanalyzer) to semi-automatically measure RNA quantity and sizes. In short, isolated RNA was labeled with a fluorescence tag and forced through a gel matrix via electrophoresis. Precise measurement of the fluorescence on the other side of the gel resulted in a *quasi*-size-concentration graph exemplarily depicted in Figure 2.2, much similar to a numerical representation of an agarose gel electrophoresis. Because of their abundance and known sizes, identification and quantification of rRNAs is straight forward. Looking at our measurements, we found that rRNAs indeed do not accumulate equally. As shown in Figure 2.7B, amounts of long rRNAs do oscillate in light-dark conditions. The concentration of 23S rRNA increased during the night from about 20 $\frac{\text{ng}}{\text{ml}}$ to roughly 60 $\frac{\text{ng}}{\text{ml}}$, corresponding to a triplication of the amount, but decreased rapidly when the light was switched on. Likewise, the slightly shorter 16S rRNA increased from 20 $\frac{\text{ng}}{\text{ml}}$ to 40 $\frac{\text{ng}}{\text{ml}}$ and also dropped back to 20 $\frac{\text{ng}}{\text{ml}}$ as soon as the day began. In contrast, the concentration of

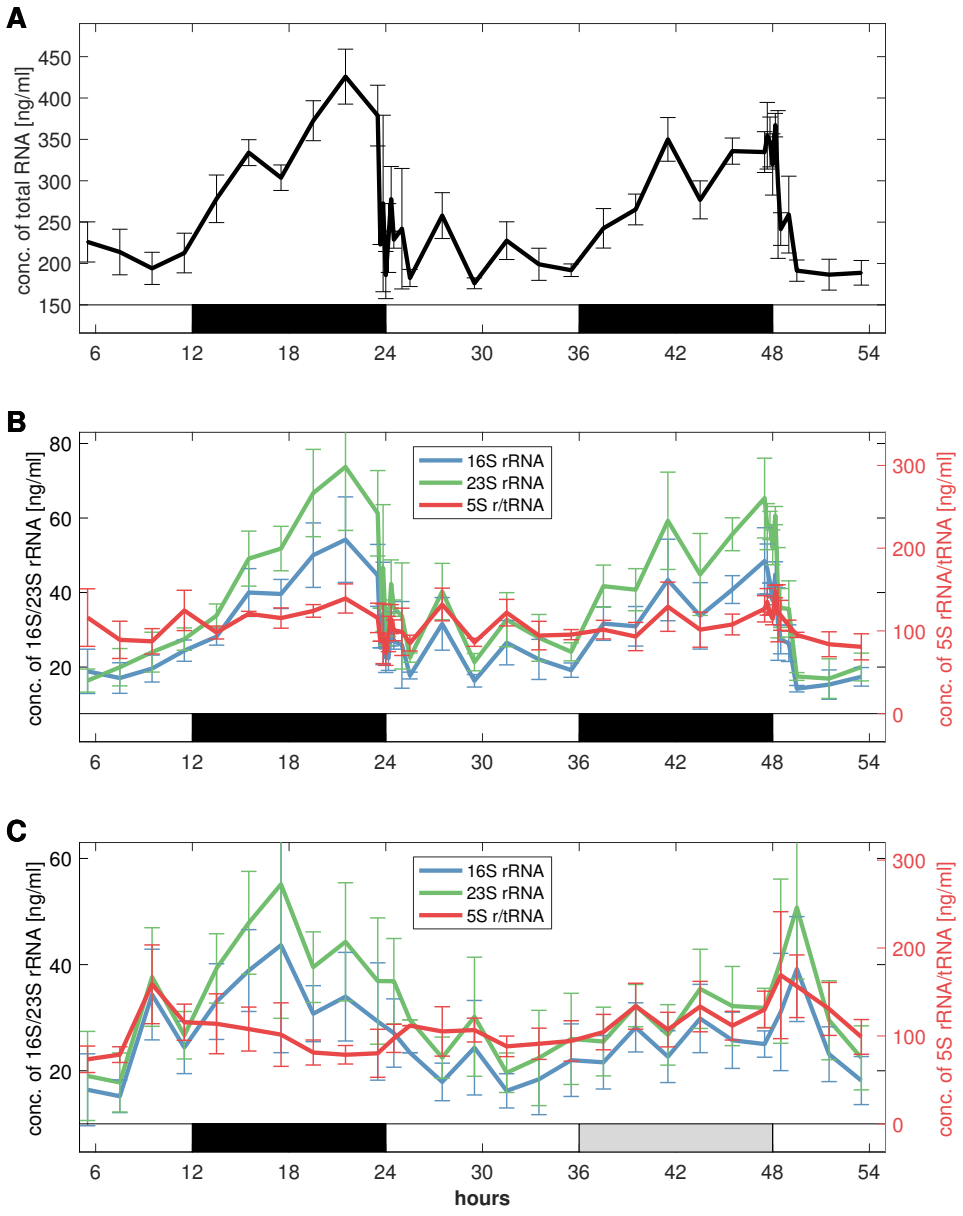


Figure 2.7.: Composition and total concentration of RNA in one milliliter culture with OD_{750}=1. **(A)** The absolute concentration of RNA shows a clear diurnal rhythm with a strong accumulation during the night and rapid decrease at the onset of light. Those changes in concentration, however, are not uniform for all RNAs **(B)**. Long 16S and 23S ribosomal rRNAs have high amplitudes, while short 5S rRNA and tRNAs remain constant throughout the time series. Furthermore, the accumulation of long RNAs does not persist during the subjective night in the LDLLL experiment and is therefore not controlled by the circadian clock **(C)**. Data points show the average of four biological replicates measured with the Agilent Bioanalyzer. Total RNA of all replicates was in addition measured with the Nanodrop™ ND-1000. Samples containing contaminated or degraded RNA were omitted. Whiskers indicate the standard error of the mean. Black boxes on the x-axis depict dark periods while the gray box in C depicts the subjective night of the LDLLL experiment. Please note the different scaling of 16S/23S rRNA on the left and 5S rRNA on the right y-axis in panels B and C.

5S rRNA and tRNAs remained equal during day and night phases at around 100 ng per milliliter culture. However, oscillations in the amount of long rRNAs is most certainly not driven by a circadian but rather a direct response to the absence of light, since the accumulation vanished when the cells were transferred to continuous light conditions as shown in Figure 2.7C.

In addition to the diurnal accumulation of long rRNAs, we did observe cleavage of the 23S rRNA. Mature 23S rRNA with a size of 2,883 nt is partially split into two fragments of sizes 500 nt and 2,400 nt (Figure 2.2). This process was previously described for the cyanobacterium *Synechococcus* sp. PCC 6301 (formerly named *Anacystis nidulans*) by Doolittle 1973. Doolittle further reported that the cleavage was stimulated by light. However, we carefully measured the concentration of each fragment independently, but could not determine significant changes in the ratio of cleaved and mature 23S rRNA during light or dark phases. In *Synechocystis*, this process is therefore most certainly triggered by a different, yet to be identified process.

2.4. Discussion

Using high-density microarray chips, three different light regimes, and a high sampling rate of one sample every two hours, we conducted the most comprehensive study on circadian gene expression in any cyanobacterium so far. We observed a tight schedule of upregulated processes in *Synechocystis* throughout a 24 hour cycle in alternating light-dark condition that was adequate for the phototrophic lifestyle of the cyanobacterium. Processes related to energy metabolism, translation, and phototrophic growth were activated right at the onset of light. Towards the end of the day, genes related to oxygenic respiration and catabolism were upregulated in anticipation of the upcoming dark phase. Sensitive processes such as transmembrane transport as well as DNA replication and repair were shifted to the night, when cells experienced less oxygenic stress. Similar schedules have been reported in previous transcriptome studies of cyanobacteria, performed under various conditions (Labiosa et al. 2006; Ito et al. 2009; Vijayan et al. 2009). However, diurnal oscillations vanished in constant conditions and we could not identify a single transcript that was controlled by a sustained circadian clock, thus showing rhythmic expression in both continuous light and continuous dark conditions. This is puzzling, since circadian rhythms were observed - albeit for only a small number of genes - in continuous light conditions (Aoki et al. 1995; Kucho et al. 2005) as well as for heterotrophic growth in continuous darkness (Aoki et al. 1997). In a more recent study, Pascal van Alphen and his colleague demonstrated persistent circadian rhythms in various growth parameters of *Synechocystis* cultivated in a continuously illuminated photobioreactor (van Alphen and Hellingwerf 2015). In an other study, we verified the physiological function of the three KaiC proteins in *Synechocystis* (Wiegard et al. 2013). Phosphorylation of KaiC1 indeed depended on the binding of KaiA, as was reported for KaiC in *Synechococcus*, suggesting a similar timing mechanism. KaiC2 and KaiC3 on the other hand, showed strong autophosphorylation activity, independent of KaiA. The same was reported for a reduced hourglass-like clock mechanisms, composed of only

KaiB and KaiC, in *Prochlorococcus marinus* (Axmann et al. 2009).

Upon close inspection of the core clock genes, we observed extensive regulation by *cis*-encoded asRNAs, resulting in low-amplitude expression of the *kai* genes. In fact, relative diurnal amplitudes for *kai* genes in *Synechocystis* are only a fraction of the circadian amplitudes found *in vivo* in *Synechococcus* under continuous light conditions (Ito et al. 2009; Vijayan et al. 2009). Considering the observations, we argue that *Synechocystis* most likely features a switchable clock system that can be altered through the regulation of asRNAs. Depending on environmental conditions, the clock may switch from an hourglass-like diurnal timekeeper to a self-sustained circadian oscillator. Hourglass timer are sufficient to structure the day by measuring the passage of time, however need a daily external cue to reset the clock. In *Synechocystis* the clock is most likely reset by the transition from the dark to the light phase and the main advantage of such a timing mechanism seems to be the preparation of the organism in anticipation of the upcoming night, towards the end of the light phase. Hourglass-like timers have not only been observed in cyanobacteria (Holtzendorff et al. 2008; Axmann et al. 2009) but in purple photosynthetic bacteria as well (Ma et al. 2016). Conducting various knock-out experiments in *Rhodopseudomonas palustris*, Ma and colleagues demonstrated a strong growth advantage of cells possessing a functional hourglass timing mechanism, compared to cells with a disrupted clock, when growing in environments with a 12 h-light-12 h-dark cycle. Temporal organization of intracellular processes by these simple diurnal timers therefore seems to be sufficient and advantageous for an optimal growth of organisms living in mostly regular environments (Johnson et al. 2017).

The presence of an hourglass-like timing mechanism in *Synechocystis* is also supported by a previous study by Anderson and McIntosh. *Synechocystis* requires a short daily light pulse to grow heterotrophically, with glucose as sole source of energy and carbon, in the dark (Anderson and McIntosh 1991). This requirement for light-activated heterotrophic growth remained enigmatic for a long time, as the light pulse is insufficient for phototrophy. Considering our observation, we argue that the daily light pulse is necessary to reset the hourglass-like timer, thereby allowing the activation of essential biological processes such as respiration, translation, and replication. A later study, however, showed that *Synechocystis* indeed can be entrained to grow heterotrophically in continuous darkness (Aoki et al. 1997), without the need of a daily light pulse, and exhibits circadian rhythms under such conditions. This observation and later studies reporting circadian rhythms, suggest that *Synechocystis* is able to activate a fully functional circadian clock if favored by the conditions. These can include heterotrophic growth (Aoki et al. 1997) but also rapid phototrophic growth in well illuminated CO_2-rich environments (van Alphen and Hellingwerf 2015). The exact conditionality of the circadian clock, however, remains unknown and requires further investigations. Silencing of the circadian clock has been reported for *Synechococcus*, which is unable to produce persistent oscillations in continuous light at temperatures below 23° C. Reactivation of the circadian clock at these low temperatures has a detrimental effect on phototrophic growth rates (Xu et al. 2013a). Multiple copies of *kaiB* and *kaiC* and the extensive regulation of circadian clock genes by multiple *cis*-antisense RNAs in *Synechocysis* might therefore be a necessity to fine-tune the circadian clock in response to various environmental

conditions experienced by this organism.

In addition to the tightly regulated diurnal schedule of gene activation, we observed strong oscillation in the amount of total RNA, which accumulated slowly throughout the night but vanished rapidly at the onset of light. This behavior seemed to be driven by the absence of light rather than controlled by a circadian clock, as the oscillation did not proceed in continuous light. Concentration of RNA is thought to be closely related to the growth rate (Kjeldgaard and Kurland 1963; Binder and Liu 1998), yet we observed virtually no increase in cell density during night phases. Moderate accumulation of RNA has been reported for *Synechococcus* sp. PCC 6301, but was attributed to a general shift of cell devision towards the light phase (Lepp and Schmidt 1998). Main contributors to the accumulation of total RNA are the highly abundant long ribosomal 16S and 23S RNAs, as rRNAs and tRNAs account for the largest share of total RNA. However, the shorter 5S rRNA and tRNAs did not show significant rhythmicity but remained at equal levels in light and dark phases. To our knowledge, this phenomenon has not been described for *Synechocystis* or any other cyanobacteria before. Thus far, we can only speculate about reason and mechanism for the rRNA oscillations. Previous research on other organisms showed massive changes in the concentration of rRNA in response to stress, yet the studies remain inconclusive. Hansen and colleagues reported for *Lactococcus lactis* a 70% decrease of rRNA in response to heat shock stress (Hansen et al. 2001). Research on the eukaryote *Karenia brevis* in contrast revealed a two to three fold increase of rRNA in response to cold stress (Jayroe 2015). However, growing our cultures in temperature-controlled 30°C and with medium light intensity ($80\frac{mol\ \text{photons}}{m^2 s}$), we could almost certainly exclude high-light or temperature induced stress. Alternative hypotheses include that changes of the cells or ribosomes during the night alter the efficiency of the RNA extraction and purification method, and that high amounts of rRNA are required to increase the number of ribosomes during the day. Yet, we found no changes in the number of ribosomes and a second RNA purification on the supernatant after the precipitation step showed no significant amounts of RNA, neither in light nor in dark samples. Both hypotheses would also not account for the constant amount of short 6S rRNA. The most intriguing hypothesis, however, is the use of long RNAs as storage compound for ribose during the night. Nightly accumulation of long rRNAs could be a key mechanism to evade the necessity to prepare cells in anticipation of an upcoming light phase, and therefore eliminate the need for a true circadian clock.

Right at the beginning of the day, RNA nucleotides can quickly be degraded by cleaving the ribose from the nucleobases through purine- and pyrimidine-specific phosphoribosyltransferases (genes: sll1430, sll0368). The ribose can then be directly fed into the Calvin-Benson cycle by converting it to ribose 5-phosphate using a ribose-phosphate pyrophosphokinase (sll0469), where it is quickly converted to Ribulose-1,5-bisphosphate, the main substrate for the CO_2-fixing enzyme RuBisCO (Xu et al. 2013b). Rapid degradation of RNA might therefore kick-start the carbon fixation of RuBisCO immediately at the onset of light, thus reducing any delay of this process that is thought of as the bottleneck of photoautotrophic growth in standard atmospheric conditions (Woodrow and Berry 1988; Tcherkez et al. 2006). The phosphoribosyltransferases involved in this pathway are either upregulated during

the night (sll1430: cluster 7) or constitutively expressed (sll0368, cluster 9). The pyrophosphokinase (sll0469) on the other hand is diurnally expressed and activated right after the transition to the light phase (cluster 2). Using long rRNAs for storage instead of shorter 5S rRNA, mRNA, or random RNA sequences might have multiple benefits. First, rRNAs can easily be synthesized and are relatively stable. Second, longer RNAs can be packed more densely and require less space. Third, rRNAs are also essential in the formation of highly abundant ribosome complexes. Fourth, rRNAs do not possess a ribosome-binding site and therefore can not be translated to proteins. Accidental translation of storage RNA is not only very energy consuming, but the synthesized protein might have decremental effects on the organism's viability. Whether our conjecture for diurnal oscillations in the amount of long rRNAs is true, can not be clarified conclusively in the scope of this thesis. Further time-series experiments, possibly involving mass spectrometry or radioactively labeled RNA, are necessary to uncover the biochemical processes of rapid RNA degradation.

Chapter 3.

Conservation and diversity in core metabolic pathways of multiple cyanobacterial organisms

3.1. Introduction

Cyanobacteria have the ability to grow photoautotrophically in highly diverse environmental conditions (Whitton 2012; Seckbach 2007). Using sun light as sole source of energy they only need small amounts of trace minerals to convert inorganic carbon dioxide into a variety of energy-rich organic compounds such as carbohydrates, proteins, and fatty acids. This puts cyanobacteria in the midst of the most promising organisms for the sustainable production of biofuels and organic material. With doubling times in the range of 24 hours, low requirements in space and nutrition, and the ability to grow in unclean or salty water (Martins et al. 2011; Markou and Georgakakis 2011), they outperform traditional energy crops such as sugar cane, maize, and rapeseed (Dismukes et al. 2008; Pienkos and Darzins 2009). Because they don't compete with food crops for arable land, they are socially preferable as well.

To grow cyanobacteria economically in large quantity and with adequate yields, metabolic engineering of the organism's metabolism is inevitable. Multiple studies are devoted to the optimization of synthesis rates for specific biochemical products such as ethanol, hydrogen, oil, and various other valuable compounds (Lai and Lan 2015; Angermayr et al. 2015; Wang et al. 2012; Parmar et al. 2011; Ducat et al. 2011; Quintana et al. 2011). However, a fundamental understanding of the cyanobacterial metabolism is key to advance the search for efficient synthesis pathways. We, therefore, compared the genomes of 16 diverse cyanobacterial strains focusing in particular on genes with assigned metabolic functions. Directing our attention to core metabolic pathways, we were able to identify genes conserved in all strains as well as genes shared by only a subset of cyanobacteria. Conserved genes imply a fundamental set of proteins coding for cellular functions indispensable for photoautotrophic growth, such as glycolysis, carbon fixation, and glycogen storage. Genes more sparsely distributed among the organisms reflect the wide functional potentials of cyanobacteria, including for example biosynthesis of valuable compounds, fixation of anorganic nitrogen, and others. The 16 strains were selected based on variety, phylogenetic distance, and genome sequence availability. They include four α-cyanobacteria, thermophilic strains, fresh and salt water strains, filamentous and single celled organisms, as well as strains capable of nitrogen fixation.

Our study builds upon previous research comparing multiple cyanobacteria. However, these studies most often focused on the detection of phylogenetic relationship and horizontal gene transfer (Raymond et al. 2002; Zhaxybayeva et al. 2006; Shi and Falkowski 2008), analyzed proteins specific for clades of cyanobacterial (Gupta et al. 2003; Gupta and Mathews 2010), or were restricted to a small set of closely related strains (Hess 2004; Bandyopadhyay et al. 2011). In contrast, our aim was a comprehensive genomic comparison of representative cyanobacterial organisms while directing our attention on conservation and diversity of core metabolic processes.

3.2. Materials and methods

3.2.1. Selection of strains

Following visual inspection of the phylogenetic tree presented in a paper by Gupta and Mathews (Gupta and Mathews 2010), we selected a set of 16 cyanobacterial strains covering a maximum of the cyanobacterial diversity. Organisms included are: *Acaryochloris marina* MBIC11017 (Aca11017), *Cyanothece* ATCC 51142 (Cyn51142), *Cyanothece* PCC 8801 (Cyn8801), *Gloeobacter violaceus* PCC 7421 (Glo7421), *Microcystis aeruginosa* NIES-843 (Mi843), *Nostoc* sp. PCC 7120 (Nos7120), *Prochlorococcus marinus* MED4 (ProMED4), *Prochlorococcus marinus* MIT 9211 (Pro9211), *Prochlorococcus marinus* MIT 9215 (Pro9215), *Synechococcus* JA-2-3B_a (SycJA23), *Synechococcus* sp. PCC 7002 (Syc7002), *Synechococcus* sp. WH7803 (Syc7803), *Synechococcus elongatus* PCC 7942 (Syc7942), *Synechocystis* sp. PPC 6803 (Syn6803), *Thermosynechococcus elongatus* BP-1 (ThermoBP1), and *Trichodesmium erythraeum* IMS101 (Trich101). General biological information of the cyanobacterial strains including preferred habitat, morphology, genome size, and G+C content are listed in the table in Appendix A. Strains considered in this study are highlighted in bold.

3.2.2. Clustering of likely orthologous genes

Chromosomal sequences of the 16 selected organisms were extracted from the NCBI GenBank database [http://www.ncbi.nlm.nih.gov/genbank] (Benson et al. 2008). To identify orthologs we performed for amino acid sequences of every gene a similarity search against every gene in all 16 genomes using BLASTp (Altschul et al. 1990). The bidirectional hit rate (BHR) between two genes a and b was computed as

$$\mathrm{BHR}_{a,b} = \left(\frac{S_{a,b}}{S_b^{bestA}}\right) \times \left(\frac{S_{b,a}}{S_a^{bestB}}\right),$$

where $S_{a,b}$ is the BLAST score of a and b with a as query, and S_b^{bestA} is the best BLAST score of b against any gene in the genome containing a (Moriya et al. 2007). The BHR ranges between zero, if the similarity of both genes is below the BLAST score cutoff (<20), and one, if a and b are the mutually best hits in their respective genomes. The BHR between pairs of genes from the same genome was limited to 0.95. Gene pairs with an BHR of 0.95 or above were classified as preliminary orthologs and subsequently merged. To avoid clusters where two genes having a

low BHR are weakly connected by a third gene, provisional groups of orthologs were clustered a second time based on their mutual BHR using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) and a cutoff at 0.75 (Sokal 1958). Cutoff values were determined through visual inspection of the cluster. The resulting clusters of likely orthologous genes (CLOGs) are composed of genes with high sequence similarity, therefore assumably coding for the same function. Every CLOG thus represents the orthologous copies of one gene across all strains. Gene annotations of any gene gathered from the GenBank file were expanded to all genes within the same CLOG.

3.2.3. Enrichment of GO annotation

For general functional classification of genes, we relied on the annotations provided by the Gene Ontology (GO) Consortium. The GO database (Harris et al. 2004) was searched for annotations for every gene of the 16 strains. Each CLOG was attributed with all GO terms annotated to any assigned gene. Enrichment of gene functions in core and unique CLOGs were calculated using Fisher's exact test and the parent-child algorithm to identify the most relevant and significant GO terms. Both methods are implemented in the TopGO package (Alexa et al. 2006) available as part of the Bioconductor software suite for R [www.bioconductor.org]. For the calculation of p-values only CLOGs with assigned GO term where taken into account.

3.2.4. Assignments of metabolic function

In addition to the broad functional GO annotations we were interested in the specific metabolic function of CLOGs. Therefore, all genes of the 16 strains considered here were matched to the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2008) and possibly labeled with the respective Enzyme Commission (EC) numbers, indicating distinct enzymatic activity. EC numbers were expanded to the according CLOGs, and CLOGs with at least one assigned EC number were considered as metabolic. CLOGs could be assigned with multiple metabolic annotations, as all EC numbers of associated genes were considered unless one number was just an incomplete form of another annotated EC number (e.g. 1.11.1.- and 1.11.1.15). Multiple EC numbers assigned to one CLOG are indicative of either multifunctional enzymes, faulty annotations, or erroneous clustering of orthologous genes. However, out of 1,851 metabolic CLOGs only 52 were assigned to more than one distinct EC number. In 18 of these cases, EC numbers differed in only the fourth serial digit, meaning that the enzymes carry out the same general reaction but act on slightly different substrates. For example NAD dependent isocitrate dehydrogenase (EC 1.1.1.41) and NADP dependent isocitrate dehydrogenase (EC 1.1.1.42). These numbers indicate a generally low rate of inconsistent annotations even without manual curation of the annotations. In total, 802 distinct EC numbers were assigned to the 1,851 metabolic CLOGs, as isoenzymes and subunits of complex enzymes are associated to different CLOGs but annotated with identical EC numbers.

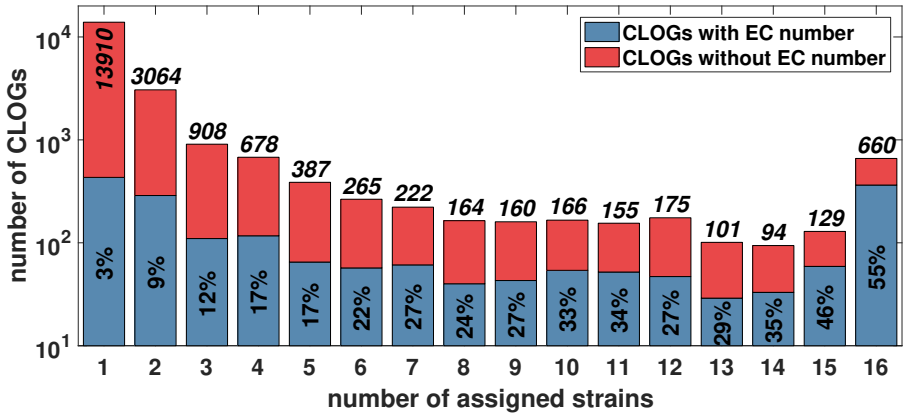


Figure 3.1.: Distribution of the number of strains associated to CLOGs. Each bar shows the number of CLOGs with (blue) and without (red) assigned EC number. Italic numbers on top of each bar indicate the total number of CLOGs, while numbers in the blue sections indicate the fraction of CLOGs with assigned specific metabolic annotations. Please note the logarithmic scale on the y-axis. The figure shows a clear decrease in the number of CLOGs with increasing number of assigned strains. While most CLOGs contain only genes from a single strain (unique), fewer genes are shared by multiple organisms. However, the number of core CLOGs shared by all 16 strains is higher than expected by this trend. The data also shows an increase in the fraction of CLOGs with a metabolic annotation in response to CLOG size.

3.3. Results

3.3.1. From core to pan-genome

The clustering of likely orthologous genes resulted in the identification of 21,238 distinct CLOGs, which can be divided into the three categories core, shared, and unique. Core CLOGs are composed of at least one gene for every strain and represent genes that are conserved in all 16 strains. Shared CLOGs contain genes from more than one but not all genomes and are therefore specific for a certain subset of strains. Unique CLOGs are single genes that have no orthologs in any other genome. The size of CLOGs, meaning the number of strains associated with at least one gene, has a clear distribution, as shown in Figure 3.1. Most of the CLOGs, roughly 65% (13,910), belong to the *unique* category with only one assigned strain. As a general rule, the number of distinct CLOGs decreases with CLOG size indicated by the number of assigned strains. However, as an exception, the frequency of core CLOGs associated with all strains is relatively high, indicating a quite large set of genes indispensable for photoautotrophic cyanobacteria. This results in a shifted U-shaped distribution of CLOG sizes typical for the analysis of shared genes. Most genes are either specific to a small set of organisms or essential in all strains, while only few genes are shared by a larger subset of the 16 strains. Similar distributions have been reported in previous studies comparing 12 strains of *Haemophilus influenzae* or a broad set of bacterial genomes (Hogg et al. 2007; Lapierre and Gogarten 2009).

Basically, the opposite is true for the relative number of genes with assigned metabolic functions. In unique CLOGs, only 3% of the genes were annotated with a specific EC number. In contrast, more than half of core CLOGs could be assigned to a metabolic function. This observation might be a direct consequence of the genomic annotation process. Biochemical functions of genes are typically analyzed in only few model organisms and subsequently assumed for homologous genes in other organisms identified by sequence similarity. Therefore, genes shared by many organisms have a higher chance of being functionally analyzed in detail compared to genes specific for only few organisms. On the other hand, our results highlight the huge untouched biochemical potentials as many unique genes will code for enzymes with strain-specific functionality awaiting thorough examination.

The previous observation not only holds true for assignments of metabolic function but for annotations in general as well. This was tested by comparing the GO annotations assigned to all CLOGs. While over 90% of core CLOGs have dedicated functional GO annotations to at least one of their genes, this number drops below 54% for shared and 22% for unique CLOGs, again indicating an inferior understanding of less common genes. Nonetheless, we were able to perform an enrichment analysis of GO terms associated to core and unique CLOGs. Mainly three biological processes were significantly enriched in unique CLOGs, meaning they were more often annotated for unique genes than one would expect by chance: regulation of gene expression (1.1×10^{-22}, GO:0010468), regulation of cellular biosynthetic process (1.3×10^{-21}, GO:0031326), and defense response (1.6×10^{-6}, GO:0006952). In contrast, enriched GO terms in core CLOGs mostly refer to "housekeeping" and metabolism-related functions such as translation (10^{-30}, GO:0006412), cellular amino acid metabolic process (3.1×10^{-28}, GO:0006520), biosynthetic process (1.4×10^{-25}, GO:0009058), cellular ketone metabolic process (1.8×10^{-22}, GO:0042180), and RNA modification (2.4×10^{-8}, GO:0009451). This enrichment analysis suggests that enzymatic functions are more likely conserved in a plethora of organisms, while the regulation of such processes is highly specific for single cyanobacterial strains, possibly adjusting common metabolic processes to individual environmental constraints. Intuitively, defensive response to biological threats is dependent on the occupied environment and cohabiting organisms, and thus also highly specific for individual strains. However, as we discussed earlier, enrichment of metabolic processes in core CLOGs might just be an artifact of inadequate annotation of metabolic functions for genes associated to only few genomes.

3.3.2. Extrapolation of core and pan-genome

So far we calculated the sizes of core and pan-genomes considering all 16 cyanobacterial strains. However, the thorough examination of gene orthology opened up the opportunity to extrapolate these results beyond the number of strains considered here. For that, we randomly selected 2 to 16 strains and recalculated the size of core and pan-genome, shown in Figure 3.2. Naturally, the size of the core genome between only two selected strains was high but reduced rapidly the more strains are considered in the comparison. Because of functional and phylogenetic relationship of cyanobacteria, this number will always asymptotically approach a value larger

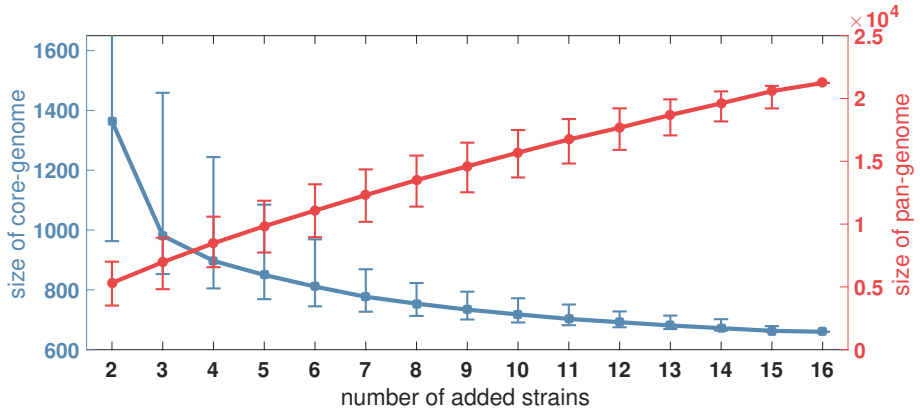


Figure 3.2.: Size of the core and pan-genome for randomly sampled genomes. Core and pan-genomes were calculated from 10,000 random samples of each 2 to 16 genomes. Size of the core genome (blue line) was calculated as the number of CLOGs shared by all selected organisms and indicated on the left y-axis. Size of the pan-genome (red line) is the number of CLOGs associated to any selected CLOG and is indicated on the right y-axis. Whiskers stretch from the 10%-quantile to the 90%-quantile. The core genome of all 16 strains comprises 660 CLOGs. Based on the trajectory, we carefully estimated a core genome size of roughly 600 CLOGs for cyanobacteria capable of oxygenic photosynthesis and photoautotrophic growth. The pan-genome of all 16 strains comprises 21,238 CLOGs.

than zero. Based on our data, we extrapolated the number of core genes that could be expected if more than 16 strains would be compared. Roughly estimated, 600 genes seemed to be shared by most photoautotrophic cyanobacteria. We emphasize that this number is only a very rough approximation, as we can not rule out rare genetic events based on our 16 strains. Genes attributed to the core genome in this comparison might be absent in a rare set of cyanobacteria. This can not be predicted here precisely because of the rarity of such events (Kislyuk et al. 2011). It is also important to understand that the core genome does not reflect the genome of a minimal photoautotrophic organism. Among other reasons, alternative biochemical mechanisms for indispensable cellular functions eliminate the involved genes from the core genome. Cyanobacteria from the genus *Prochlorococcus* for example use divinyl chlorophyll as main light-harvesting pigment, while most other cyanobacteria use monovinyl chlorophyll for the same purpose (Chisholm et al. 1992). Although chlorophyll is indispensable for cyanobacterial photoautotrophic growth, some of the enzymes required for biosynthesis of these pigments are therefore not included in the core genome.

Size of the pan-genome by definition is much larger than the core genome. Considering all 16 strains, it encompasses more than 21,000 CLOGs and increases with the number of included genomes, showing almost no flattening of the curve (Figure 3.2, red line). Extrapolating the trend in our data, we predict an open pan-genome for cyanobacteria with approximately 500 novel CLOGs for every new genome added to the comparison. Including more strains in the analysis will therefore most likely result in the discovery of a high number of yet unknown orthologous genes. The

question of whether pan-genomes of bacterial clades are open or closed, has been controversially discussed in the literature. Finite pan-genomes were reported for *Haemophilus influenzae* and *Streptococcus pneumoniae* (Hogg et al. 2007; Donati et al. 2010), whereas infinite size of the pan-genome was predicted for the cyanobacterial genus *Prochlorococcus* and *Streptococcus agalactiae* (Tettelin et al. 2005; Kettler et al. 2007). These contradicting results may reflect differences in the evolutionary history of the selected organisms as well as in the sizes of the ecological niches, and thus the availability of genetic variations.

3.3.3. Genome sizes of cyanobacterial strains

Going beyond the pan-genome analysis, we were particularly interested in the diversity of the 16 cyanobacterial strains. In accordance with their variations in morphology, lifestyle, and habitats, cyanobacteria show a high diversity in their genome size, ranging from 1.69 million bp for the streamlined genome of *Prochlorococcus marinus* MIT 9211 up to 6.5 million bp for *Acaryochloris marina* MBIC11017. This is reflected in the number of associated CLOGs - depicted in Figure 3.3 - which ranges accordingly between 1,800 and 5,600. In large parts, this variability can be attributed to a high diversity in unique genes, as for example more than a third of the CLOGs of *Acaryochloris*, *Microcystis*, and *Gloeobacter* are unique. In contrast, the relatively small genomes of *Thermosynechococcus* and *Prochlorococcus* contain only a small fraction of unique CLOGs.

Compared with the majority of non-metabolic clusters, CLOGs assigned to at least one EC number, depicted in Figure 3.3 below the continuous line, show a very different distribution of corresponding strains. Most notably, the number of metabolic CLOGs is remarkably similar across all strains. Core CLOGs make up between one third and up to 60% of the clusters assigned to a specific metabolic function. In contrast, less than 3% of the unique CLOGs are annotated with an EC number. As discussed earlier, this observation can likely be explained by strong conservation of metabolic functions and bias in the annotation of EC numbers. Our GO term analysis of core and unique CLOGs suggests that metabolism-related genes are significantly conserved in all genomes, while regulation- and environment-specific processes are more diverse. However, functional annotation of genes is mostly accomplished through assessment of sequence similarities, thereby creating an inherent bias towards genes shared by multiple organisms.

3.3.4. Diversity of the cyanobacterial metabolism pathways

Our primary topic of interest in this study was the understanding of conservation and diversity of the central metabolism in cyanobacteria. Thorough knowledge on essential as well as dispensable pathways might indicate new approaches for genetic modification and biosynthetic capabilities of cyanobacteria. The central metabolism can be divided into five core pathways, namely glycolysis, the Calvin-Benson-Bassham (CBB) cycle, the pentose phosphate pathway, pyruvate metabolism, and the tricarboxylic acid (TCA) cycle. In addition, we investigated the pathways for biosynthesis of the three storage compounds glycogen, polyhydroxybutyrate (PHB), and

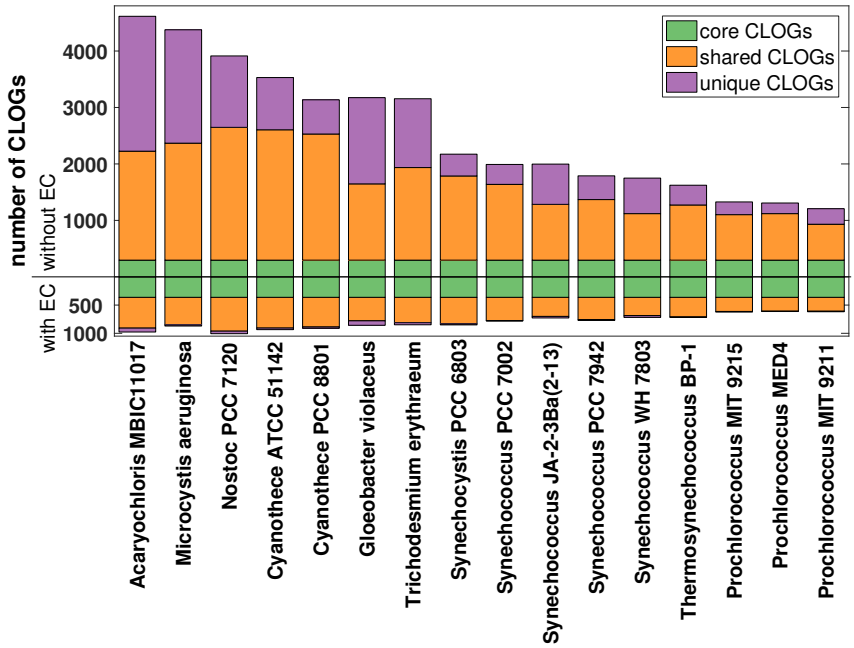


Figure 3.3.: Number of core (green), shared (orange), and unique (purple) CLOGs assigned to each of the 16 cyanobacterial strains. The bars are separated into CLOGs with assigned EC number below the continuous line and others above. Strains are sorted by genome size. CLOGs with an EC number have a high ratio of core genes while shared and unique CLOGs most often have no EC number assigned. Interestingly, unique CLOGs have a higher variability between the strains than shared CLOGs.

cyanophycin. A conclusive map of all pathways indicating conserved and partially shared enzymes is depicted in Figure 3.4. Detailed association of cyanobacterial strains to particular pathways is presented in Figure 3.5. In the following, all pathways will be thoroughly discussed and differences in the metabolic capabilities of organisms will be compared to the literature.

3.3.5. Glycolysis and pentose phosphate pathway

Glycolysis (Figure 3.4, highlighted in green) is the anaerobic conversion of glucose, glucose 1-phosphate, and other related monosaccharides into two molecules pyruvate, thereby releasing energy in form of two ATP and two NADPH. This pathway is central for the energy metabolism throughout all domains of live (Romano and Conway 1996; Müller et al. 2012). Since most participating enzymes are bidirectional, this pathway also plays an essential role in glucogenesis, the synthesis of glycogen (Romano and Conway 1996). Therefore, it is not surprising to find most of the required enzymes conserved in all cyanobacterial strains. The only exceptions being the phosphofructokinase (PFK, EC 2.7.1.11) and fructose-1,6-bisphosphatase (FBP, EC 3.1.3.11) that convert fructose 6-phosphate (F6P) to fructose-1,6-bisphosphate (FBP) and vice versa. These enzymes are missing in the α-cyanobacteria *Prochlorococcus* and *Synechococcus* sp. WH 7803. Flux balance analysis of the metabolic net-

work of *Synechocystis* PCC 6803 revealed that these reaction might be dispensable for biomass synthesis (Knoop et al. 2010). In a recent paper, Chen and colleagues confirmed the absence of PFK in these strains as well as other cyanobacteria (Chen et al. 2016). The affected strains, however, utilize the alternative Entner-Doudoroff (ED) pathway - not shown in Figure 3.4 - which makes use of the two enzymes phosphogluconate dehydratase (EC 4.2.1.12) and KDPG aldolase (EC 4.1.2.14) to convert 6-Phosphogluconate (6PG) into glyceraldehyde 3-phosphate (GAP). This pathway has lower costs for enzymes but also yields less ATP, which seems to be acceptable for photoautotrophic organisms living in open oceans and being limited on nutrients rather than ATP (Chen et al. 2016).

The pentose phosphate pathway is an alternative series of reactions for the anabolism of glucose. It can be separated into the oxidative phase converting glucose 6-phosphate to ribulose 5-phosphate, marked light blue in Figure 3.4, and the non-oxidative phase transforming ribulose 5-phosphate to fructose 6-phosphate, erythrose 4-phosphate, and other sugars. Reactions of the second phase overlap with the CBB cycle. Overall, the pentose phosphate pathway can generate more reducing equivalents in the form of NADPH but less ATP compared to glycolysis. In cyanobacteria and chloroplasts of plants, this pathway is considered to be important for the synthesis of NADPH as well as provision of ribulose-5-phosphate (Ru5P) required for synthesis of nucleotides and replenishing the CBB cycle (Turner and Turner 1980). It is therefore no surprise that the enzymes are conserved in all 16 cyanobacterial strains as well.

3.3.6. Carbon fixation in cyanobacteria

Fixation of atmospheric carbon dioxide (CO_2) in cyanobacteria is achieved via the CBB cycle, also known as reductive pentose phosphate cycle or C3 cycle - marked purple in Figure 3.4 (Bassham et al. 1954). This cyclic process can be separated into three stages. Inorganic CO_2 is transferred by the enzyme Ribulose-1,5-bisphosphat-carboxylase/-oxygenase (RuBisCO, RBCO) onto a total of three molecules ribulose 1,5-bisphosphate (RuBP) generating six molecules of 3-phosphoglycerate (PG3). In the second reductive phase, each molecule of PG3 is reduced to glyceraldehyde 3-phosphate (GAP). Finally, in the regenerative phase five molecules of GAP are used for a series of transformations forming again three molecules of RuBP, thus completing the cycle. In summary, on each iteration of the full cycle three molecules of CO_2 are fixed and converted to one molecule of PG3. This PG3 can subsequently be stored as glycogen using the glycolysis pathway or converted to a plethora of metabolites using TCA cycle and other pathways. The CBB cycle is the predominant pathway for the fixation of atmospheric CO_2 into carbonaceous metabolites in photoautotrophic organisms such as cyanobacteria. It is therefore no surprise to locate the required enzymes in all 16 investigated strains. In fact, because the core enzyme RuBisCO has a relatively low turnover rate of only 1,000-2,000 molecules per minute, it is the most abundant protein in cyanobacteria as well as plants (Tabita 2004) and probably the most common protein on earth (Ellis 1979; Raven 2013). To boost enzyme efficiency in cyanobacteria, RuBisCO is encapsulated in proteinaceous micro compartments with internally increased CO_2 concentration called carboxysomes

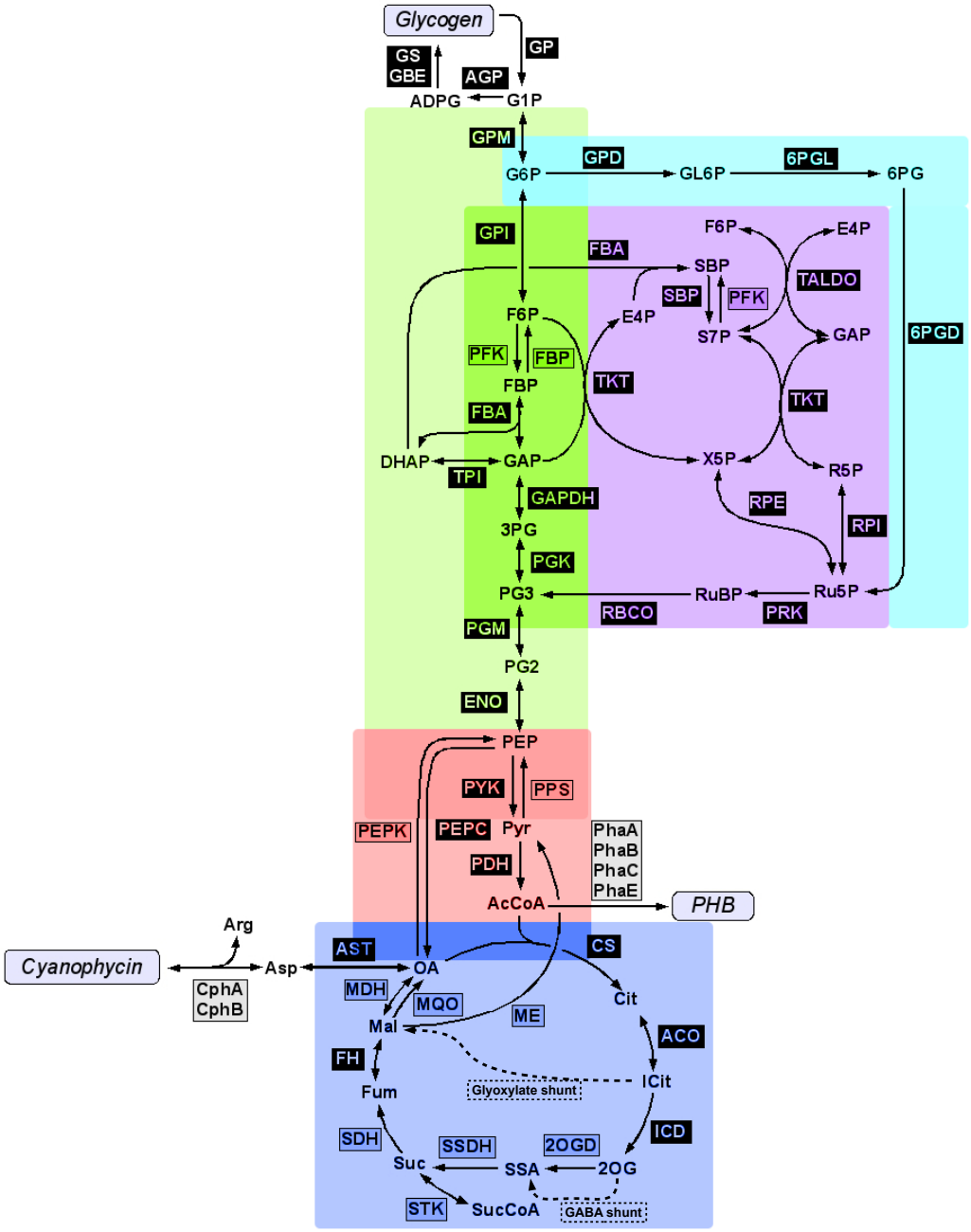


Figure 3.4.: Storage and core metabolic pathways in cyanobacteria. Enzymes in black boxes are associated with all 16 strains; enzymes in gray/colored boxes are shared by only a subset of cyanobacteria. The core metabolism is subdivided into glycolysis (green), Calvin-Benson-Bassham cycle (purple), pentose phosphate pathway (cyan), pyruvate metabolism (red), and tricarboxylic acid (TCA) cycle (blue). Storage metabolism includes pathways for synthesis of glycogen, cyanophycin, and polyhydroxybutyrate (PHB). Most core pathways and glycogen synthesis are very much conserved in all 16 cyanobacteria, whereas the TCA cycle and synthesis of cyanophycin and PHB are more fragmented. Abbreviations: 2OGD: 2-oxoglutarate decarboxylase (EC 4.1.1.71), 6PGD: phosphogluconate dehydrogenase (EC 1.1.1.44), 6PGL: Phosphogluconolactonase (EC 3.1.1.31), ACO: Aconitase (EC 4.2.1.3),

Figure 3.4 cont.: AGP: ADP glucose pyrophosphorylase (EC 2.7.7.27), AST: aspartate transaminase (EC 2.6.1.1), CphA: Cyanophycin synthetase (EC 6.3.2.29/30), CphB: Cyanophycinase (EC 3.4.15.6), CS: Citrate synthase (EC 2.3.3.1), ENO: Enolase (EC 4.2.1.11), FBA: Fructose-bisphosphate aldolase (EC 4.1.213), FBP: Fructose-1,6-bisphosphatase (EC 3.1.3.11), FH: Fumarate hydratase (EC 4.2.1.2), GAD: Glutamate decarboxylase (EC 4.1.1.15), GAPDH: Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.12/59), GBE: Glycogen branching enzyme (EC 2.4.1.18), GP: Glycogen phosphorylase (EC 2.4.1.1), GPD: G6P dehydrogenase (EC 1.1.1.49), GPI: Glucose-6-phosphate isomerase (EC 5.3.1.9), GPM: Glucose phosphomutase (EC 5.4.2.2), GS: Glycogen synthase (EC 2.4.1.21), ICD: Isocitrate dehydrogenase (EC 1.1.1.41/1.1.1.42), MDH: Malate dehydrogenase (EC 1.1.1.37), ME: Malic enzyme (EC 1.1.1.38), MQO: Malate:Quinone oxidoreductase (EC 1.1.5.4), PEPC: PEP carboxylase (EC 4.1.1.31), PEPK: PEP carboxykinase (EC 4.1.1.49), PDH: Pyruvate dehydrogenase (EC 1.2.4.1), PFK: Phosphofructokinase (EC 2.7.1.11), PGK: Phosphoglycerate kinase (EC 2.7.2.3), PGM: Phosphoglycerate mutase (EC 5.4.2.1), PhaA: PHA-specific β-ketothiolase/Acetyl-CoA acetyltransferase (EC 2.3.1.9), PhaB: PHA-specific acetoacetyl-CoA reductase (EC 1.1.1.36), PhaC/E: Poly(3-hydroxyalkanoate) synthase (EC 2.3.1.-), PPS: PEP synthetase (EC 2.7.9.2), PRK: Phosphoribulokinase (EC 2.7.1.19), PYK: Pyruvate kinase (EC 2.7.1.40), RBCO: Ribulose 1,5-bisphosphate carboxylase/oxygenase (EC 4.1.1.39), RPE: Ribulose-5-P 3-epimerase (EC 5.1.3.1), RPI: Ribose-5-P isomerase (EC 5.3.1.6), SBP: Fructose-1,6-/Sedoheptulose-1,7-bisphosphatase (EC 3.1.3.37), SDH: Succinate dehydrogenase (EC 1.3.99.1), SSDH: succinate-semialdehyde dehydrogenase (EC 1.2.1.16) STK: Succinate thiokinase (EC 6.2.1.5), TALDO: Transaldolase (EC 2.2.1.2), TKT: Transketolase (EC 2.2.1.1), and TPI: Triosephosphate isomerase (EC 5.1.3.1). Parts of this figure were adapted from Beck et al. 2012.

(Rae et al. 2013). Interestingly, two evolutionary distinct forms of carboxysomes are known, enclosing two different types of RuBisCO. RuBisCO form-1A is used by α-cyanobacteria, mostly marine *Synechococcus* and *Prochlorococcus* strains and contained in α-carboxysomes. RuBisCO form-1B, on the other hand, is encapsulated by β-carboxysomes and found in all other (β-)cyanobacteria as well as in higher plants (Rae et al. 2013). Plants, in contrast, evolved different C_4 and crassulacean acid metabolism (CAM) CO_2-concentrating mechanisms (Keeley and Rundel 2003).

3.3.7. Pyruvate metabolism and TCA cycle

In contrast to the previously discussed pathways, metabolic processes involving pyruvate are rather fragmented in the 16 studied cyanobacterial strains (Figure 3.4, marked red). While the anaplerotic reactions converting phosphoenolpyruvate (PEP) to oxaloacetat (OA) and citrate (Cit), thus replenishing the TCA cycle, are conserved in all strains, the cataplerotic reactions, extracting metabolic intermediates from the TCA cycle, are not (Figure 3.5). Genes for the PEP carboxykinase (PEPK, EC 4.1.1.49) converting oxaloacetat to PEP can only be found in *Microcystis aeruginosa*

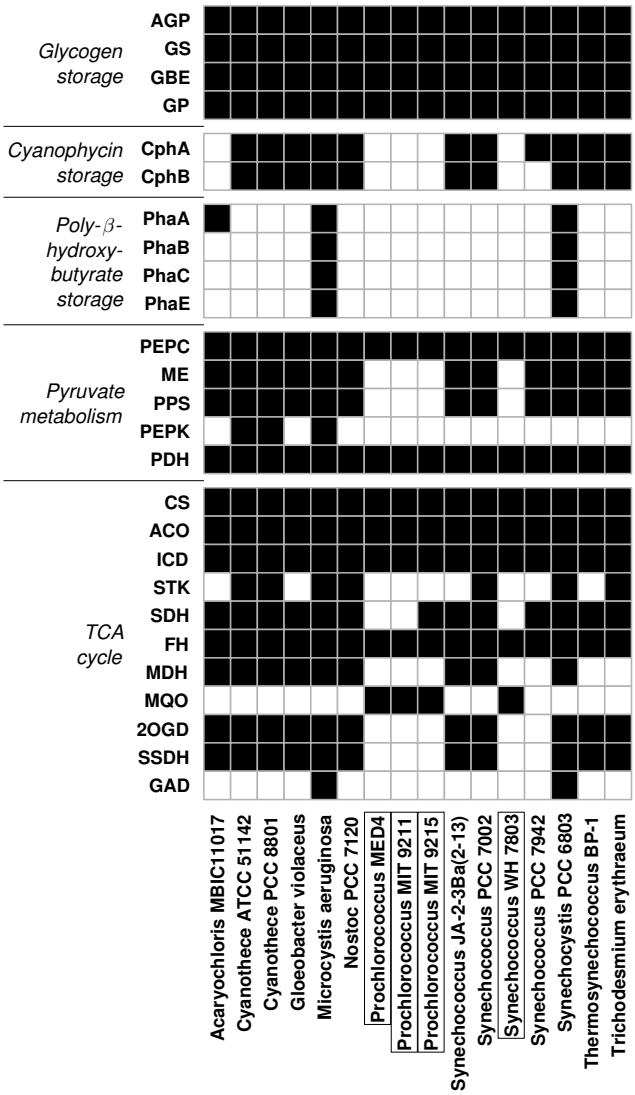


Figure 3.5.: Association of core metabolic processes with cyanobacterial genomes. This figure lists the occurrence of enzymes involved in five metabolic processes in all 16 cyanobacterial strains. This includes pathways for biosynthesis of the storage compounds glycogen, cyanophycin, and PHB as well as the pyruvate metabolisms and the TCA cycle. Each line represents one specific enzyme. Black boxes indicate if this enzymatic function could be assigned to at least one gene of the strains indicated on the x-axis. α-cyanobacteria are highlighted by black lines enclosing the species names. Abbreviations of enzymes are identical to Figure 3.4.

NIES-843 as well as both strains of *Cyanothece* (ATCC 51142 and PCC 8801). The malic enzyme (ME, EC 1.1.1.38) and PEP synthetase (PPS, EC 2.7.9.2) catalysing the convertion of malate (Mal) to pyruvate (Pyr) and subsequently to PEP are missing in the α-cyanopbacterial strains *Synechococcus* sp. WH 7803 and *Prochlorococcus*. Both enzymes play an essential role in the C_4 and CAM CO_2-concentrating pathway

of plants (Edwards and Huber 2014; Ting 1985) but the function in cyanobacteria remains enigmatic. Bricker and colleagues measured a reduced growth rate in constant light for a malic enzyme knockout mutant of *Synechocystis* PCC 6803. However, this growth deficit vanished if cells were supplemented with glucose or under diurnal lighting conditions with 12h light and 12h dark (Bricker et al. 2004). In a recent paper Yoshikawa et al. measured identical growth rates for a similar mutant with a malic enzyme knockout and the wild type *Synechocystis* under constant light conditions. However, the production rate of ethanol was slightly diminished (Yoshikawa et al. 2015b). Overall the findings indicate that in α-cyanobacterial strains, the cataplerotic reactions might be dispensable due to reduced growth rates in their nutrient-poor open ocean habitats. Alternatively they are accomplished by other, yet to be identified enzymes.

The TCA cycle (also citric acid or Krebs cycle, marked blue in Figure 3.4) is a central metabolic pathway that is - at least partially - present in most aerobic organisms throughout all domains of life (Huynen et al. 1999; Schnarrenberger and Martin 2002). Upon each completion of the cycle, one acetyl here provided as acetyl-coenzyme A (AcCoA) is fully reduced to each two molecules CO_2 and water, thereby generating reducing agents in the form of NADH and FAD, which are fed into the respiratory chain to generate ATP. The TCA cycle therefore constitutes the final step in the complete breakdown of sugars and provides over 90% of the reduction equivalents generated from carbohydrate catabolism (Buchanan et al. 2015). Additionally, the pathway provides essential precursors for the synthesis of multiple cellular components, most notably various amino acids generated from 2-oxoglutarate (2OG) and oxaloacetate (OA) (Riccardi et al. 1989; Zhang and Bryant 2011). In cyanobacteria, the cycle is highly fragmented as indicated in Figure 3.5. Most notably, the enzyme complex 2-oxoglutarate dehydrogenase, catalyzing the conversion of 2-oxoglutarate to Succinyl-CoA (SucCoA) is missing in all cyanobacteria (Pearce et al. 1969; Smith et al. 1967), thereby interrupting cyclic flow. In 2011, Zhang and Bryant reported the discovery of two novel enzymes (2-oxoglutarate decarboxylase (2OGD) and succinate-semialdehyde dehydrogenase (SSDH)) in *Synechococcus* sp. PCC 7002 that convert 2-oxoglutarate to succinate semialdehyde (SSA) and subsequently to succinate (Suc), thus bypassing the absent 2-oxoglutarate dehydrogenase and restoring the TCA cycle (Zhang and Bryant 2011). However, this SSA bypass is absent in the α-cyanobacterial *Prochlorococcus* and *Synechococcus* sp. WH 7803 strains as well as in *Synechococcus elongatus* PCC 7942. Interestingly, the right hand side of TCA cycle converting oxaloacetate and acetyl-CoA to 2-oxoglutarate is highly conserved in all cyanobacteria. Because the succinate thiokinase (STK) is not needed for cyclic flow, the according gene is missing in nine of the 16 strains. Also the succinate dehydrogenase (SDH) is absent in all α-cyanobacteria except for *Prochlorococcus marinus* PCC 9215. Beginning at fumarate (Fum), which is an byproduct of phototrophic growth (Knoop et al. 2010), the TCA cycle is again somewhat conserved in all studied strains. However, while most strains utilize malic enzyme (ME) or the bidirectional malate dehydrogenase (MDH) for conversion of malate (Mal), α-cyanobacteria rely on the multifunctional but unidirectional Malate:Quinone oxidoreductase (MQO) to close the cycle. Overall, circular flow through the TCA cycle is feasible for most studied cyanobacteria, with the exception of α-cyanobacterial

Prochlorococcus and *Synechococcus* sp. WH 7803 strains and possibly *Synechococcus elongatus* PCC 7942. Yet, reactions required for the synthesis and degradation of essential metabolites like fumarate, oxaloacetate, and 2-oxoglutarate are conserved in all organisms.

Two other pathways capable of completing the TCA cycle, namely the glyoxylate shunt and the GABA shunt (shown in dashed lines in Figure 3.4), are discussed in literature (Steinhauser et al. 2012). The glyoxylate shunt consists of two enzymes, isocitrate lyase (EC 4.1.3.1) converting isocytrate to glyoxylate and malate synthase (EC 2.3.3.9) converting the glyoxylate to malate. This pathway has been described for two strains of the genus *Cyanothece* (Bandyopadhyay et al. 2011; Gründel et al. 2017). Yet we could not identify either enzyme in any of the 16 cyanobacterial genomes. The second pathway converts glutamate, synthesized from 2-oxoglutarate to gamma-Aminobutyric acid (GABA), utilizing a glutamate decarboxylase (GDC, EC 4.1.1.15). GABA is subsequently converted to succinate semialdehyde (EC 2.6.1.19). The central enzyme GDC could only be identified in *Microcystis aeruginosa* NIES-843 and *Synechocystis* sp. PCC 6803 (Figure 3.5).

Overall, our data suggest that the TCA cycle of α-cyanobacteria as well as *Synechococcus elongatus* PCC 7942 is interrupted and can not generate cyclic flow. As a result, these organisms are obligate phototrophs unable to efficiently utilize external carbohydrates. For α-cyanobacteria living in the open ocean, extracellular sugar is a scarce commodity. It therefore is not energetically efficient to maintain the enzymes necessary for a functional TCA cycle. Low demands of energy for basic cellular maintenance during the night can probably be covered by glycolysis and the pentose phosphate pathway. Obligate phototrophy was also reported for *Synechococcus elongatus* PCC 7942 (Zhang et al. 1998), however the exact cause remained enigmatic thus far. Interestingly, parts of the TCA cycle leading to 2-oxoglutarate and fumarate are highly conserved in all strains, as these intermediates are essential for biosynthesis of various amino acids and other vital metabolites.

3.3.8. Biosynthesis of storage compounds

Cyanobacteria rely on sun light as their primary source of energy. To maintain basal cellular functions during the night, part of the energy generated by photosynthesis is stored as polymeric compounds. In our analysis of the central metabolisms, we therefore included the pathways for synthesis and breakdown of three storage compounds with high relevance for cyanobacteria, namely glycogen, cyanophycin, and poly-β-hydroxybutyrate (PHB) (Figures 3.4 and 3.5). These storage compounds are of particular interest for biotechnological applications, as they are rich in energy, carbon, as well as other nutrients, and are produced naturally in high concentrations.

Glycogen is a branched polysaccharide made from glucose that serves as storage for carbon and energy. In its chemical composition and function, it is highly similar to starch that is used for energy storage by plants. In cyanobacteria, glycogen is accumulated through coalescing glucose 1-phosphate and ATP to ADP-glucose catalyzed by ADP glucose pyrophosphorylase (AGP), which in turn is polymerized to a macromolecule of glycogen by glycogen synthase (GS). For a compact storage structure, the glycogen chain is irregularly branched by the glycogen branching en-

zyme (GBE) (Kromkamp 1987). Sequential breakdown of glycogen is carried out by the enzyme glycogen phosphorylase (GP) removing one glucose monomer from the glycogen chain and releasing it in the form of glucose 1-phosphate. Glycogen or similar α-polyglucan structures are synthesized by all cyanobacteria (Nakamura et al. 2005), and it is therefore no surprise that the related enzymes are conserved in all 16 strains studied.

Cyanophycin is an amino acid polymer comprised in equal parts of arginine and aspartate, and serves in cyanobacteria as a storage compound for nitrogen (Kromkamp 1987). This polymer attracted interest of the chemical industry as a biodegradable dispersant and organic source for polyaspartic acid, which has multiple technological applications as superabsorber, thickner, and in water treatment (Oppermann-Sanio and Steinbüchel 2002). Cyanophycin is aggregated through alternating elongation of the cyanophycin chain with arginine and aspartate, conducted by the single enzyme cyanophycin synthetase (CphA). Degradation of cyanophycin is carried out by the enzyme cyanophycinase (CphB). Synthesis of the polymer is reported for various cyanobacterial strains as well as multiple non-cyanobacterial organisms (Krehenbrink et al. 2002). In our study, neither CphA nor CphB could be identified in *Acaryochloris marina* MBIC11017 or any studied α-cyanobacteria. These strains are therefore unable to synthesize cyanophycin. The same is likely true for *Synechococcus elongatus* PCC 7942, which lacks an orthologous gene for CphB, and whose gene for CphA has relatively low sequence similarity to CphA of *Synechocystis*.

The biopolyester PHB is synthesized and retained in granules by multiple microbial organisms including some cyanobacteria (Steinbüchel et al. 1998; Asada et al. 1999). It serves as a storage for excess carbon and energy under stress conditions and emerged as a sustainable, non-toxic, and biodegradable alternative for petroleum-derived polyesters (Philip et al. 2007; Balaji et al. 2013). Synthesis of PHB begins with the condensation of two Acetyl-CoA molecules to Acetoacety-CoA (PHA-specific β-ketothiolase PhaA), which is in turn reduced to 3-Hydroxybutanoyl-CoA (PhaB). The latter is polymerized to PHB by polyhydroxyalkanoate synthase (PhaC and PhaE). The required pathway could be identified in only two of the 16 strains, namely *Microcystis aeruginosa* NIES-843 and *Synechocystis* sp. PCC 6803. *Acaryochloris marina* MBIC11017 is associated with the CLOG for PhaA but lacks the other enzymes required for PHB synthesis. Certainly, this gene is an ortholog for a non-PHB-specific ketothiolase. *Microcystis aeruginosa* and *Synechocystis* therefore remain the two most interesting organisms when it comes to sustainable production of polyesters from PHB.

3.4. Discussion

Increased availability of fully sequenced bacterial genomes offers novel approaches to understand microbial diversity. In this study, we established a new method for a genome-wide genetic comparison of multiple photoautotrophic cyanobacteria with the goal to understand the diversity of cyanobacterial metabolism. Cyanobacteria inhabit a huge variety of environments yet share a fundamental photoautotrophic lifestyle. This is also reflected in the distribution of orthologous genes. Of the 21,238

CLOGs identified in this study, about 65% consist of only a single gene with no ortholog in any other considered strain. About 3% of the CLOGs, however, are shared by the 16 considered strains. Careful inspection of the data suggests a core set of roughly 600 genes shared by all photoautotrophic cyanobacteria. Yet, we emphasize that such extrapolations must be met with caution and that the core genome does not represent a minimal gene set for photoautotrophic organisms. In contrast, we found no indication of a closed cyanobacterial pan-genome. Thus, expanding the genome-wide comparison to a broader set of cyanobacterial strains would inevitably reveal novel genetic functions and metabolic pathways.

Core CLOGs represent an indispensable set of cellular functions while unique CLOGs indicate specific adjustments of single strains to their environmental conditions. Both sets therefore differ significantly with respect to their functional annotations. Unique CLOGs are typically related to processes of adaptation such as regulation of gene expression and cellular defense. Core CLOGs on the other hand are mostly annotated with housekeeping functions including translation and metabolic processes. Consequently, we observed a huge bias in the annotation with specific metabolic functions as well. While 55% of the core CLOGs could be assigned with a distinct EC number, the same was true for only 3% of the unique CLOGs. Thorough investigation of five central metabolic pathways revealed rather strong conservation of the genes involved in the pentose phosphate pathway, CBB cycle, and glycolysis. Our analysis results in reliable assignments of metabolic capabilities, as we predicted an impaired glycolysis in α-cyanobacteria due to the absence of phosphofructokinase, which was confirmed in a study by Chen and colleagues (Chen et al. 2016). Two other pathways involved in the complete break-down of carbohydrates, namely pyruvate metabolism and the TCA cycle, are highly fragmented across the strains. As a result, oxidation of acetyl-CoA is impaired in α-cyanobacteria and *Synechococcus elongatus* PCC 7942. Fragments of the pathways, however, are obligatory for the biosynthesis of essential amino acids and therefore conserved in all strains including α-cyanobacteria. In addition, we examined the distribution of pathways related to anabolism and catabolism for three storage compounds of particular interest for biotechnological application. Biosynthesis of glycogen is conserved in cyanobacteria, while pathways for PHB and cyanophycin could only be observed in a subset of strains.

While focusing on the core metabolism, several conclusions can be drawn from our analysis. First, cyanobacteria are highly diverse even in central metabolic pathways such as the TCA cycle. Especially strains of the α-cyanobacterial clade show exceptional adaptations to their nutritionally poor habitat, resulting in impairment of pyruvate metabolism and TCA cycle as well as the inability to synthesize cyanophycin. Second, thorough genetic comparison can be used to identify specific traits in microbial organisms with possible relevance for various biotechnological applications, e.g. biosynthesis of cyanophycin and the biopolymer PHB. Third, concurring assignment of organisms to multiple metabolic CLOGs can reveal the underlying pathway structures. Most of the genes involved in cyanophycin synthesis, PHB synthesis, or the succinate semialdehyde bypath co-occur in a subset of strains as none of the related enzymes alone is sufficient for the metabolic process und will be eliminated from the genomes throughout evolution. Fourth, extrapolation of the

pan-genomes suggest that the presented method should be expanded to a wider set of cyanobacterial organisms. Increase in strain diversity would not only result in a more precise estimate for the size of core and pan-genome but also allow examination of rare, non-central pathways.

As a consequence, this analysis was followed by a second study presented in the next Chapter 4 after genome sequences of novel cyanobacterial strains were made available by Shih and others (Shih et al. 2013). Extending the genome-wide comparison to 77 cyanobacterial strains allows us to systematically analyze co-occurrence, identify diversity in multiple non-central pathways, reconstruct the metabolism, and mathematically simulate biosynthesis in all examined strains. In this respect, we consider our analysis also as a first step towards an automated network reconstruction.

Chapter 4.

Analysis of co-occurring genes improves understanding of metabolic diversity in cyanobacteria

By refining the method described in the previous Chapter 3, we were able to expand the analysis of orthologous genes to a group of 78 prokaryotes by including the genomes of 77 diverse cyanobacterial and one *Escherichia coli* strain. Widen the set of organisms not only allowed for a refined analysis of the core- and pan-metabolims. But by analyzing the occurrence of genes we were able to identify groups that co-occurred in most similar subsets of organisms. Such co-occurring genes show significant similarities in their biological function although they are not necessarily co-located on the respective genomes. To facilitate the analysis of co-occurring genes beyond the examples discussed in this thesis, we created a mathematical toolbox called "CyanoCLOG SimilarityViewer". Furthermore, our method lead to a thorough understanding of the enzymatic capabilities of the cyanobacterial strains, allowing the automated reconstruction of metabolic models for each organism. These models were systematically examined using flux balance analysis (FBA) and successfully compared to published metabolomic studies.

4.1. Introduction

Identifying the biological function of an unknown gene is a crucial but difficult task. It usually involves the comparison of sequence or predicted structural elements with according databases. If this does not result in similarities to genes with an established annotation, labor-intensive biochemical studies are often necessary. An alternative bioinformatic approach is to identify functionally linked proteins by searching for genes that occur in the same set of organisms. This method is based on the assumption that proteins involved in shared pathways or complex structures are likely to be correlatively evolved as well. If both genes are necessary for a vital biological process, they would co-occur in organisms utilizing this process. Otherwise, they would slowly be eradicated during the course of evolution. This idea, also known as phylogenetic profiling, was put forward as soon as the genomic sequence was determined for a crucial number of organisms (Marcotte et al. 1999; Pellegrini et al. 1999). In their study, Pellegrini and colleagues compared the absence/presence of homologues of more than 4,000 genes in sixteen bacterial genomes. Profiles with not more than three bits difference (the presence of a homolog differs in maximal three

of the sixteen genomes) were labeled as neighbors. Genes involved in basal processes like translation, transcription, but also metabolic processes like glycolysis and purine biosynthesis showed a higher rate of co-occurrence than groups of randomly selected genes.

With soaring availability of genomic sequences, analysis of co-occurrence was accordingly enhanced by increasing the number of genomes (Kim and Price 2011) and domains of life. Thus, subsequent studies not only covered bacterial organisms but also included genomes from archaea (Enault et al. 2003; Zhou et al. 2006; Cokus et al. 2007) as well as eukaryotes (Vert 2002; Date and Marcotte 2003; Wu et al. 2003; Sun et al. 2005; Snitkin et al. 2006; Barker et al. 2007; Jothi et al. 2007; Škunca and Dessimoz 2015). With larger numbers of genomes, more sophisticated methods emerged. After all, using the Hamming distance as measure for profile similarity (Pellegrini et al. 1999) is computationally simple and intuitive but tends to overestimate the number of correlated pairs. Core genes (present in all organisms) for example are by definition neighbors with a distance of zero but might not have a direct functional relation. Unique genes (present in only one organism) also have low distances to each other while they might not even be present in the same organism. Therefore, comparison of profiles in succeeding studies is based on Pearson's correlation coefficient (Enault et al. 2003; Wu et al. 2003), hypergeometric distribution (Cokus et al. 2007; Kharchenko et al. 2006; Ruano-Rubio et al. 2009; Simonsen et al. 2012), mutual information (Date and Marcotte 2003; Sun et al. 2005; Snitkin et al. 2006; Jothi et al. 2007), maximum likelihood models (Barker et al. 2007), or machine learning algorithms (Škunca and Dessimoz 2015). Another factor for the performance of phylogenetic profiling methods is the construction of the profiles. Most approaches use binary profiles, assuming the presence of a homologous gene, if the similarity of a gene sequence exceeds a defined threshold, and absence otherwise (Wu et al. 2003; Cokus et al. 2007; Kharchenko et al. 2006; Barker et al. 2007; Jothi et al. 2007; Ruano-Rubio et al. 2009; Simonsen et al. 2012; Škunca and Dessimoz 2015). Other studies seek to preserve information of the characteristic of the homology by discretizing the similarity score into predefined bins (Date and Marcotte 2003; Snitkin et al. 2006) or using continuous scores in the profiles (Enault et al. 2003). Comparison of the approaches, however, led to no conclusive results. Methods using continuous scores performed better in some sets of organisms, while discretized profiles were superior in others (Snitkin et al. 2006).

Today thousands of prokaryotic and hundreds of eukaryotic genomes have been fully sequenced. However, multiple studies have shown that including more and more genomes in the analysis of co-occurring genes does not improve the results but has a rather decremental effect on the algorithm's performance. Škunca and Dessimoz showed that including more than 100 genomes had almost no effect on the results, while inflating the computational effort (Škunca and Dessimoz 2015). Studies by Snitkin and Jothi found that including more than a few eukaryotic genomes reduced the accuracy of detecting correlated genes, even when searching for pathways in the eukaryote *Saccharomyces cerevisiae* (Snitkin et al. 2006; Jothi et al. 2007). Beyond a certain number of organisms, adding more genomes only increases the noise while adding basically no novel information on genes involved in a joint biological process. Incorporating genomes of parasites with often incomplete pathways will even

impede the detection of correlated genes. In addition, eukaryotes tend to have a more structured genome. Functionally related genes are frequently fused together into a single open reading frame, further complicating the construction of phylogenetic profiles (Apic et al. 2001). Various studies attempt to optimize the set of incorporated genomes to maximize the predictive power, mostly by maximizing the phylogenetic distance of selected organisms (Sun et al. 2005; Cokus et al. 2007; Simonsen et al. 2012; Škunca and Dessimoz 2015), while selecting a random set is almost as effective (Škunca and Dessimoz 2015). However, including very diverse organisms does not always improve the accuracy of the detection of co-occurrence. In fact, Jothi and colleagues argued that the evolutionary history of the genomes included in a study should correspond to that of the expected co-occurring processes (Jothi et al. 2007).

Comparative studies have shown that carefully selecting the method for quantifying co-occurrence, e.g. mutual information or hypergeometric distribution, has only little effect on the results. However, taking the evolutionary history of the compared profiles into account has a huge impact on the accuracy of the predictions (Barker et al. 2007; Cokus et al. 2007). The rationale here is the following: assuming the phylogenetic tree of all organisms, genes that co-occur in multiple branches of the tree are more likely to be truly co-evolved than genes that co-occur in a single clade of the tree. To obtain the former, multiple events of concerted horizontal gene transfer or gene deletion are necessary, while the latter might be a lineage-specific effect. Cokus and colleagues provided a simple solution by listing the organisms according to their evolutionary distance and accounting for the number of *runs* - uninterrupted stretches of co-occurrence in the list (Cokus et al. 2007).

Considering all the published methods, we opted for an customized workflow to analyze the co-occurrence in cyanobacteria, picking and adapting elements from multiple studies. We developed an algorithm based on the adjusted mutual information (AMI) and integrated the evolutionary distance of occurrences using a method similar to that of Cokus et al. (Cokus et al. 2007). The AMI is a variation of the mutual information used in previous studies but accounting for lopsided profiles (Vinh et al. 2010). While the mutual information is by definition low for sparsely or densely packed profiles, the AMI is a robust similarity measure in these cases as well. Evolutionary distance within profiles was taken into account by generating a phylogenetic tree and computing the consisteny scores, basically a calculation of the minimal number of genetic events (deletion, horizontal gene transfer) necessary to obtain a given profile (Kluge and Farris 1969). In the attempt to specifically identify pathways in cyanobacteria, we included 77 cyanobacterial and only one other prokaryotic organism in our workflow. In contrast to most other studies, we were not only interested in co-occurring gene pairs, but also identified modules of multiple correlated genes to detect complex biological structures and pathways. In the scope of this thesis, we can only discuss a number of modules while all co-occurring pairs of genes identified with our method and parameter set are listed in Appendix C. To allow readers and researchers simple access to our data set, while also enabling the adjustment of the parameter set to their needs, we created a simple and intuitive computer program called *CyanoCLOG SimilarityViewer* available at [http://sourceforge.net/p/similarityviewer/]. This tool facilitates the quick and easy identification of homologous and co-occurring genes for any given gene or gene cluster.

The methods used to identify homologous and functionally correlated genes resulted in a comprehensive understanding of the biological capabilities of the investigated cyanobacteria. Using that knowledge, we were able to reconstruct the metabolic network of all 78 organisms. Genome-scale metabolic network reconstructions provide a comprehensive compendium of biochemical reactions taking place in a living organism and are a key approach in determining the genotype-phenotype relationship. Such models have been successfully used for a number of applications including directing of metabolic engineering, quantifying species relationships, and identifying optimal growth conditions (Oberhardt et al. 2009; Kim et al. 2012; Monk et al. 2013; Bordbar et al. 2014; Magnúsdóttir et al. 2017). Although manual reconstruction is a complex process (Thiele and Palsson 2010), manually curated networks have been published for a variety of organisms including several cyanobacteria (Baroukh et al. 2015; Cogne et al. 2003; Montagud et al. 2010; Saha et al. 2012; Vu et al. 2012; Knoop et al. 2013; Triana et al. 2014; Yoshikawa et al. 2015a; Malatinszky et al. 2017). Multi-strain reconstructions allowing for a comparative genome analysis have been described for multiple genera including *Lactococcus* (Notebaart et al. 2006), *Escherichia coli* (Monk et al. 2013), *Cyanothece* (Mueller et al. 2013), as well as the human gut microbiome (Magnúsdóttir et al. 2017). In addition, several platforms aiming at semi-automated, high-throughput generation and analysis of metabolic models are available. Most notably are SEED (Henry et al. 2010), MetRxn (Kumar et al. 2012), RAVEN Agren et al. 2013), and BioCyc (Caspi et al. 2014). To account for genomic incompleteness and spurious annotations, these methods usually rely on manual curation or an automated gap-filling process to complete metabolic pathways (Vitkin and Shlomi 2012).

We used flux balance analysis (FBA), a constrain-based modeling approach assuming constant concentrations for every internal metabolite to verify all 78 metabolic reconstructions and compare their complexity. Even though we refrained from using any gap-filling process, these models show high consistency with available biochemical literature and therefore provide decent indication of the organism's biosynthetic capabilities.

4.2. Materials and methods

4.2.1. Acquisition of genomic data

We searched the NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome/) for cyanobacterial entries (January 17th, 2015) selecting all strains fully sequenced and assembled at chromosome level. We further included all associated plasmid sequences, as well as the recently annotated *Escherichia coli* O111:H (hereinafter referred to as E. coli). In total 78 chromosomes and 136 plasmids were sourced from the NCBI Genome database and are listed with their GenBank accession number in Appendix C. General information of all strains is provided in Appendix A, including among others morphology, preferred habitat, and genome size. A phylogenetic tree was constructed by extracting the 16S ribosomal RNA sequences of all genomes. Pairwise distances were calculated using the distance model by Jukes and Cantor (Jukes and Cantor 1969) and the BLOSUM62 scoring matrix. The tree was con-

structed with the *seqlinkage* function by MATLAB using the standard parameter. As expected, the only non-photosynthetic organism E. coli appeared as outgroup.

4.2.2. Clustering of likely orthologous genes

Identification of orthologous genes was done similar to the method introduced in the previous Chapter 3. Comparing the protein-coding genes of all 78 organisms with each other, we again computed the bidirectional hit rate (BHR) as

$$BHR = (\frac{S_{a,b}}{S_b^{bestA}}) \times (\frac{S_{b,a}}{S_a^{bestB}}), \tag{4.1}$$

where $S_{a,b}$ is the BLASTp score of a versus b and S_b^{bestA} is the best score of b against any gene in strain A (which includes a). Gene pairs with a BHR greater or equal 0.95 were grouped together. In a second step, genes in each group were clustered according to their mutual BLAST score using the UPGMA (unweighted pair group method with arithmetic mean) and a cut-off of 20. Genes in the resulting cluster of likely orthologous genes (CLOGs) were assumed to be orthologous, therefore sharing similar functions.

4.2.3. Computing modules of co-occurring CLOGs

Each CLOG is composed of genes occurring in a specific set of organisms. To identify CLOGs whose genes co-occur in similar subsets, we used a two-step process. For all 2.06×10^8 pairs of CLOGs neither unique (genes in only one organism) nor core (genes in all organisms), a right-sided Fisher's exact test was performed. P-values were corrected for multiple testing using the method by Benjamini and Yekutieli (Benjamini and Yekutieli 2001). Excepting a false discovery rate of 0.01, the critical p-value was at about 1.43×10^-6. For all significantly correlated pairs of CLOG i and j, their similarity $S(i, j)$ was computed as:

$$S(i, j) = AMI(i, j) \times (1 - CI(i \cap j, t)). \tag{4.2}$$

In this formula, $AMI(i, j)$ denotes the adjusted mutual information between the profiles of species participating in the CLOGs i and j. AMI is a modification of the mutual information normalizing for the entropy of the variables and is particularly well suited for lopsided frequencies (Vinh et al. 2010). The AMI ranges between zero for uncorrelated and one for fully correlated pairs, independent of the number of genes in each CLOG. The consistency index $CI(i \cap j, t)$ measures the consistency of the set of strains participating in both CLOGs i and j with the 16S rRNA phylogenetic tree t (Kluge and Farris 1969). It is defined as the number of changes of an observation divided by the minimal number of changes required to fit a given tree. For binary CLOGs, it ranges between one, if the genes can only be observed in a monophyletic group, and $1/N - 1$ - where N is the number of organisms - for maximal disseminated genes. E. coli and cyanobacterium UCYN-A were not considered for the calculation of the CI. CLOGs were grouped into modules by constructing an undirected graph with

nodes representing the CLOGs and the weight of edges $W(i, j)$ being the similarity score cut-off at 0.65:

$$W(i, j) = \max(0, S(i, j) - 0.65). \tag{4.3}$$

We used a heuristic and parameter-free algorithm for the detection of community structures in large networks by Blondel and colleagues to identify the modules (Blondel et al. 2008).

Anti-correlation between two CLOGs was quantified with the same method but using a left-sided Fisher's exact test. Correction for multiple testing yielded a critical p-value of $4.42 \times 10^{-}7$. Due to the nature of anti-correlated CLOGs, neither the consistency index nor modules can be computed for such pairs. Please note that the AMI is symmetric in regard to the observation and remains positive for anti-correlated CLOGs with plus one for fully contrary pairs.

4.2.4. Genomic adjacency

The adjacency score (AS) reflects the genomic adjacency or co-localization of genes within one module. For each organism, co-occurring genes were sorted according to their position on the DNA and the AS was computed as:

$$\mathrm{AS} = \frac{\sum_{i=2}^{n} \begin{cases} 1, & \text{if gene}_{i-1} \text{ and gene}_i \text{ in close proximity} \\ 0, & \text{otherwise} \end{cases}}{n - 1}, \tag{4.4}$$

where n is the number of genes within one module. Genes are in close proximity if they are located on the same chromosome or plasmid and separated by less than ten open reading frames in between. The adjacency score ranges between zero, if all genes are separated, and one, if they all are in close proximity. An average AS (aAS) is then computed for each module as the mean of the AS of all strains with at least two genes in the CLOGs comprising this module.

4.2.5. Annotation of CLOGs

CLOGs are labeled with the most common annotation of all associated genes taken from the corresponding NCBI GenBank files. Generic annotations such as *hypotetical*, *conserved*, *predicted protein*, and *unknown* were omitted if other annotations were identified for at least one gene. Metabolic functions of CLOGs were identified using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2008). For every organism in this study, with exception of *Calothrix* sp. 336/3, *Nodularia spumigena* CCY9414, and *Synechococcus* sp. WH 8109, we used the KEGG REST API to gather the Enzyme Commission (EC) numbers for all metabolic genes (latest update: March 6th, 2017). In addition to the KEGG data we included all enzyme numbers from the hand-curated metabolic model for *Synechocystis* sp. PCC 6803 by Knoop et al. (Knoop et al. 2013). Because of orthology we assumed identical metabolic function for all genes in a given CLOG.

4.2.6. Automated metabolic network reconstruction

Metabolic networks were constructed for all strains based on the EC numbers annotated for associated CLOGs. For each EC number, the KEGG database provides a set of explicit reactions with unique identifier for participating metabolites. The atomic formula of metabolites was also downloaded from KEGG when provided. Each reaction was checked for conservation of mass by calculating differences in the number of atoms of each kind in the substrates and products. We used symbolic computation to account for molecules with an undefined size (e.g. RNA of length n). Unspecified residues were regarded as separate elements. Reactions with an unbalanced number of any element except hydrogen were discarded. The atomic formulas did not include charges, the reactions therefore could not be checked for a balanced number of electrons.

The KEGG database stores no information on the physiologically feasible directionality of reaction. Wherever possible, directions were adopted from the Recon 2 database (version 2.04) (Thiele et al. 2013) and the hand-curated model by Knoop et al. (Knoop et al. 2013). All reactions that included the metabolite CO_2 were fixed to the direction of CO_2 consumption - unless explicitly examined by the hand-curated model - to prohibit unfeasible carbon uptake. Four reactions (R01283, R01289, R01290, R03532), theoretically allowing for unfeasible production of methionine and free gain of energy, were also fixed to the physiologically practicable directions.

In addition to the strain-specific networks comprising all reactions assigned to the strains' genes, we constructed a pan-metabolic network including all reactions assigned to any strain. We also identified 104 balanced reactions marked as *spontaneous* by KEGG, which included at least one metabolite from the pan-network. These spontaneous reactions were added to all networks. In addition, we added the bidirectional interconversion of water and carbon dioxide to bicarbonate and protons to every network:

$$H_2O + CO_2 \leftrightarrow H^+ + HCO_3^-.$$

This essential reaction is facilitated by carbonic anhydrases, which could not be identified in all cyanobacteria (Badger and Price 2003). However, this reaction can occur spontaneously, although with a relatively low rate (Badger and Price 1994).

We used FBA to simulate each metabolic network independently (Orth et al. 2010; Steuer et al. 2012). Their reactions were transferred to a stoichiometric matrix S, and the IBM CPLEX Optimizer (version 12.6.1) was used to solve the linear equation $Sv = 0$ for the flux vector v. Free uptake of 19 essential anorganic substrates (water, orthophosphate, protons, ammonia, sulfate, hydrogen sulfide, sulfite, magnesium cations, chlor ions, HCl, cobalt cations, hydrofluoric acid, selenic acid, HBr, Fe(II) ions, mercury cation, molybdate) and carbon dioxide as sole source of carbon was permitted to enable the influx of all required chemical elements. The flux of CO_2 uptake was bound to 100, while all other reactions had broader boundaries of -10,000 (if reverse direction was permitted, zero otherwise) and +10,000 (if forward direction was permitted, zero otherwise). Due to their lack of a functional Calvin-Benson-Bassham cycle, E. coli and cyanobacterium UCYN-A are unable to fix anorganic carbon dioxide and were therefore provided with a free uptake of α-D-Glucose instead. The flux of this reaction was limited to 16.6 equalizing the amount

of available carbon. All other networks were able to convert CO_2 to α-D-Glucose, thus allowing the import of glucose is not benefiting the simulations of E. coli and UCYN-A. Because all reactions are stoichiometrically balanced with the exception of import and export reactions, the amount of elements other than hydrogen is bound by the import fluxes. Elements can not be generated via infeasible cycles. This was tested for carbon by calculating fluxes in absence of CO_2 and α-D-Glucose, when no carbonaceous metabolites could be produced by any model. In cyanobacteria, carbon assimilation is mainly carried out by the enzyme Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) (Raven 1991, 2009). This was confirmed in our models by simulations with deactivated RuBisCO, when again no carbonaceous metabolite could be produced. Numeric feasibility of the linear model was ensured by adding free export reactions for all metabolites.

To simulate the synthesis of fatty acid intermediates, we had to modify the according export reactions. During synthesis, fatty acids are typically attached to an acetyl carrier protein (acp) (Volpe and Vagelos 1976). Since translation was not included in the models, we optimized for a combined hydrolyzing/export reaction

$$\text{fatty-acid-[acp]} + H_2O \rightarrow \text{acp}$$

in order to restore the carrier protein.

Biosynthetic capabilities of each network were tested by consecutive simulations maximizing the export reaction of each metabolite. Metabolites were considered to be producible, if their outfluxes exceeded 0.01, which exceeds the CPLEX solvers imprecision by orders of magnitude. The synthesis rate of each metabolite was computed as the fraction of carbon converted into that compound. For that, we multiplied the flux with the number of carbon atoms per molecule of the metabolite and divided by 100, which is the maximal carbon uptake rate. The rate of carbon-free metabolites was determined as fraction of the maximal flux computed for the same metabolite in the pan network. A total of 881 reactions could be synthesized by the pan-network.

4.3. Results

The cyanobacterial pan-genome revisited

We analyzed the orthology of genes from 78 prokaryotic organisms using a method similar to the one described in the previous Chapter 3. The genomes comprised all fully sequenced and assembled cyanobacterial genomes from the NCBI database and *Escherichia coli* O111:H (E. coli) for reference. A phylogenetic tree of all 78 organisms based on similarity of 16S rRNA is depicted in Figure 4.1.

Orthologous genes were identified based on sequence similarity using bidirectional BLASTp and grouped into clusters of likely orthologous genes, hereinafter denoted as CLOGs. Based on the number of participating organisms, we distinguish between unique CLOGs, identified only in a single strain, shared CLOGs, present in one or more but not in all strains, and core CLOGs, present in all strains with the possible exception of E. coli and cyanobacterium UCYN-A. Due to their unique properties, the cyanobacterium UCYN-A, an endosymbiont with a highly reduced genome (Zehr

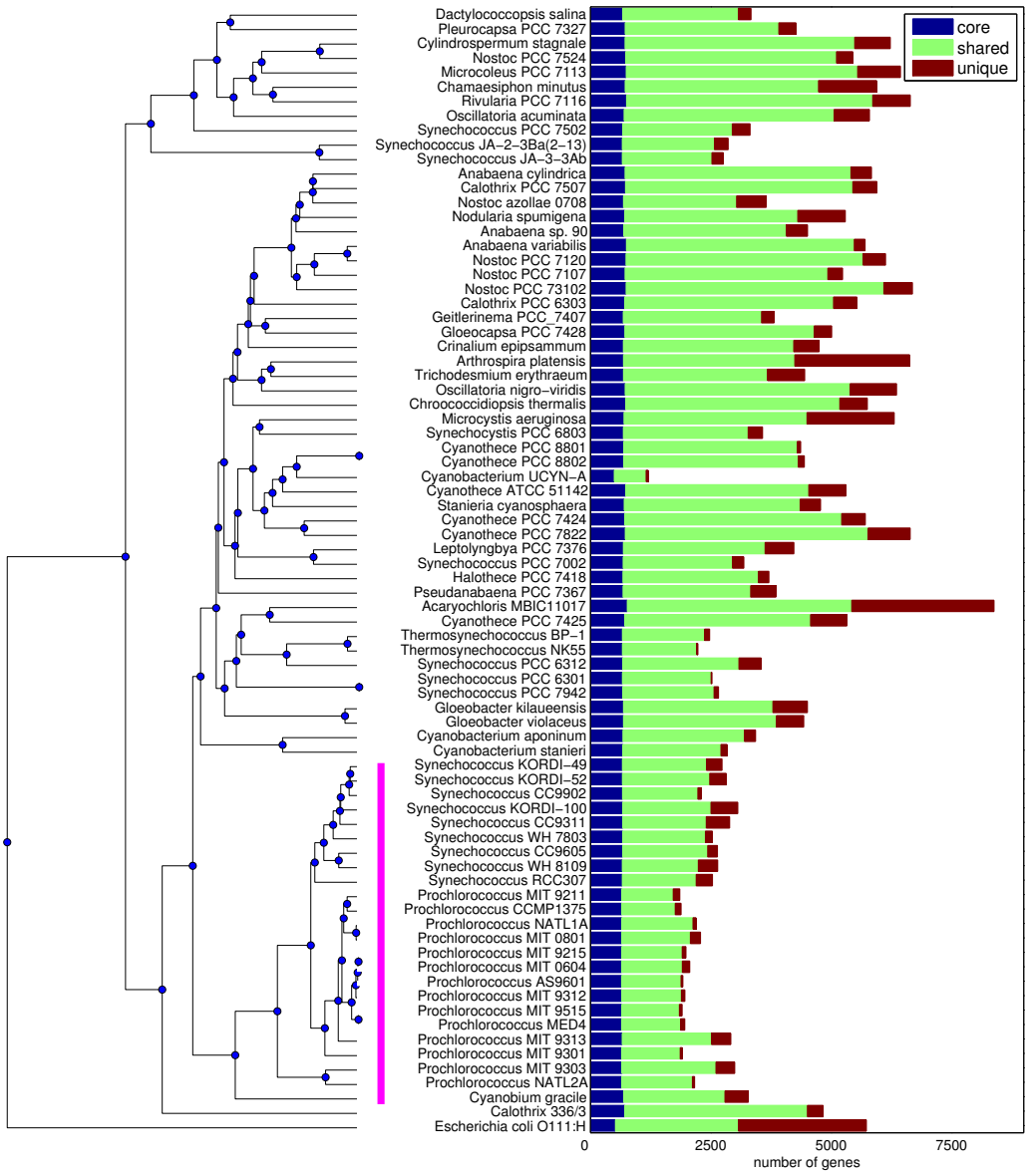


Figure 4.1.: Phylogenetic tree of all organisms. The phylogenetic tree on the left shows the lineage of all 78 bacteria based on their 16S ribosomal RNA. Pairwise distances were calculated using the method by Jukes and Cantor (Jukes and Cantor 1969) with the BLOSUM62 scoring matrix. The tree was constructed using the *seqlinkage* function of MATLAB with standard parameters. *Escherichia coli*, the only non-photosynthetic organism naturally appeared as outgroup. The vertical magenta bar marks the clade of α-cyanobacteria. The right side shows size and composition of the respective strains. The size of the bars indicates the total number of genes in the genomes divided into core (blue), shared (green), and unique genes (red). Core genes are defined as genes shared by all strains but being optional in the reduced cyanobacterium UCYN-A and *Escherichia coli*. The absolute number of core genes per strain can vary due to gene duplications.

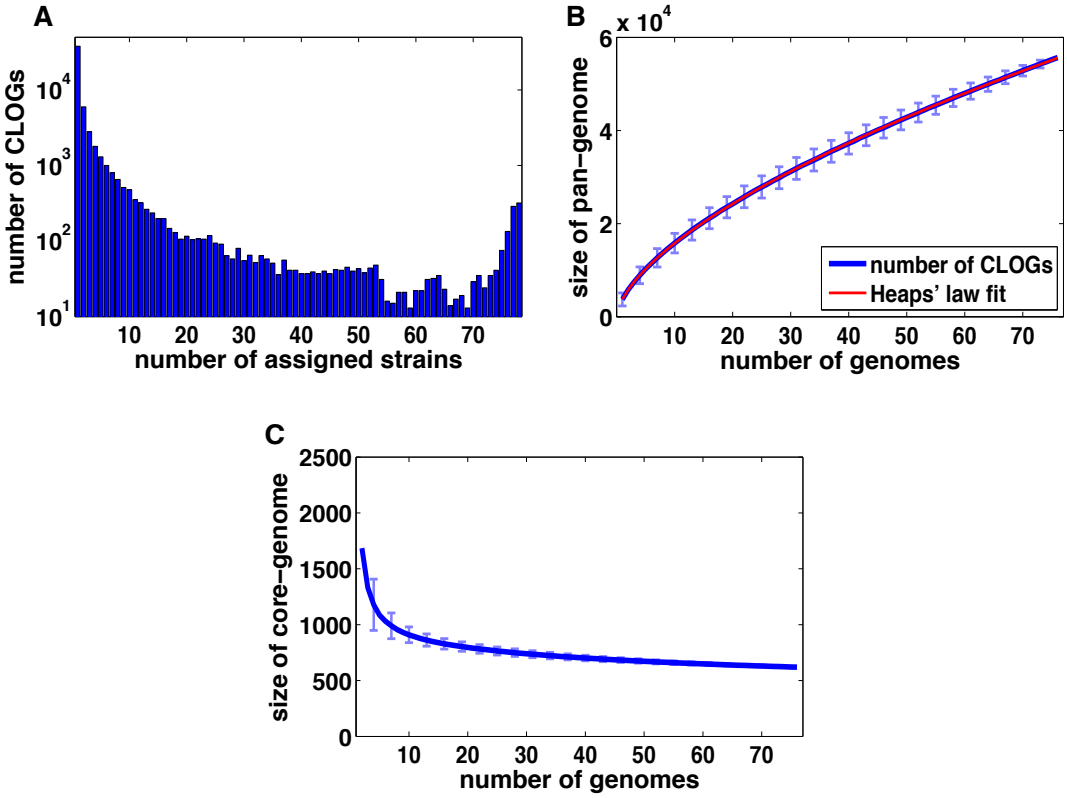


Figure 4.2.: The cyanobacterial pan- and core-genome. **A**: Distribution of CLOGs as a function of the number of assigned strains. **B**: Size of the pan-genome estimated for an increasing number of strains. The blue line indicates the mean size of the pan-genome, error bars indicate the standard deviation of 10^4 randomly sampled subsets of strains. The red line shows a least squares fit of the power law $p \approx Ng$ (Heaps' law), with p denoting the size of pan-genome and N the number of genomes. The estimated exponent $g = 0.62$ indicates an open pan-genome. **C**: Size of the cyanobacterial core-genome estimated for an increasing number of strains. The blue line indicates the mean size of the core-genome, whereas error bars indicate the standard deviation of 10^4 randomly sampled subsets of strains. The estimates of pan- and core-genome do not include genomes of E. coli and cyanobacterium UCYN-A.

et al. 2008), and the non-phototrophic E. coli were exempt from the definition of core CLOGs. We identified a total of 58,740 CLOGs consisting of 621 core, 20,005 shared, and 38,114 unique CLOGs. As depicted on the right side of Figure 4.1, strains with larger genomes tend to be associated with more shared CLOGs. The number of unique CLOGs on the other hand is also dependent on the phylogenetic distance to its nearest neighbors. The distribution of the number of strains associated with each CLOG is shown in Figure 4.2A.

The properties of the core- and pan-genome are in accordance with our findings for 16 cyanobacterial genomes (Chapter 3) as well as previous published studies (Simm et al. 2015). The core-genome constitutes between 7.4% (*Acaryochloris ma-*

rina MBIC11017) and up to 33.5% (*Prochlorococcus marinus* str. MIT 9211) of all CLOGs in a given genome. Evaluation of the pan-genome for subsets of strains reveals an increase in the size of the pan-genome with growing number of strains. The increase can be described by Heaps' law with a positive exponent of 0.62, suggesting an open pan-genome and approximately 450 genes with no clear similarity to any known protein for each genome newly sequenced in the near future (Tettelin et al. 2008). This number is in good agreement with a recent study by Shih and colleagues, identifying 21,107 novel genes by de-novo sequencing of 54 cyanobacterial strains, an average of 390 genes without known homolog per organism (Shih et al. 2013). Extrapolation of the core-genomes (Figure 4.2C) suggests a set of 500 to 600 genes shared by all free-living, photoautotrophic cyanobacteria. However, we again note that such interpretations should be met with caution, since limited data can not reflect rare events (Kislyuk et al. 2011).

Annotations of CLOGs have been assigned based on the most common annotations gathered from the GenBank files of their constituent genes. Unsurprisingly, core CLOGs exhibit high coverage with functional annotations, only 32 out of 621 CLOGs (about 5%) are annotated with generic terms such as hypothetical and conserved protein. The annotations of core CLOGs are enriched in fundamental categories such as transcription, translation, DNA replication, but also cellular metabolism, which again is in accordance with the previous study of 16 cyanobacteria as well as the literature (Mulkidjanian et al. 2006; Simm et al. 2015). In contrast, the rate of meaningful annotation is significantly lower for shared and unique CLOGs with 44% (8,853 of 20,005) and 82% (31,132 of 38,114), respectively, annotated as hypothetical, predicted or unknown proteins. Annotation of genes is often unspecific or varying in the exact wording and automated comparison of differing annotation is no easy task. For example, in one CLOG we found some genes annotated as "Photosystem II DII subunit", others as "Photosystem II protein D2". Although different in the wording, both terms obviously describe the same protein. In our data, we identified putatively inconsistent annotations for roughly 45% of the CLOGs comprising multiple annotated genes. However, manual inspection of 1,000 randomly selected CLOGs revealed that for only 2.5% of these, diverging annotations could not be identified as coinciding at a first glance. Differences in annotation can be a sign of incorrect clustering or erroneous annotation, but can also indicate moonlighting proteins - proteins with multiple functions (Jeffery 1999). Overall, the results give us high confidence in correct annotations for the vast majority of CLOGs.

In this work we focused on the cyanobacterial pan-metabolism. We therefore matched all genes against the KEGG database (Kanehisa et al. 2008) to identify CLOGs associated with enzymatic reactions represented by EC numbers. 2,361 CLOGs consist of at least one gene with assigned enzymatic function and thus could be related to metabolism. A total of 2,301 distinct reactions could be identified. Enzymes (and hence CLOGs) may catalyze multiple reactions, and multiple enzymes (and hence CLOGs) may catalyze the same reaction. Again, we found high enrichment of metabolic functions in core CLOGs. About 52% were associated with one or more metabolic reactions. On the contrary, this ratio is considerably lower for shared and unique CLOGs. Only 8.3% and 1.1%, respectively, could be linked to a metabolic function. Due to the high number of shared CLOGs, however, metabolic

functionality is primarily encoded in the shared genome. 1,839 of the 2,301 unique reactions are associated with at least one shared CLOG.

4.3.1. Co-occurring CLOGs indicate functional relationships

Each CLOG comprises genes occurring in a distinctive set of organisms. We hypothesize that co-occurrence of CLOGs is indicative of a putative functional relationship. That is, genes are highly likely to occur in the same subset of organisms, if their functions are mutually dependent on each other. Performing a right-tailed Fisher's exact test with multiple testing correction and an accepted false discovery rate of 0.01, we identified 581,741 (out of more than 1.7×10^9) pairs of CLOGs whose occurrences are significantly correlated. By definition, co-occurrence only relates to shared CLOGs. Core and unique CLOGs are therefore not considered in the following analysis. Co-occurrence of two CLOGs was quantified as described in methods by calculating a similarity index (SI) based on the adjusted mutual information corrected for the phylogenetic diversity of the assigned organisms using the consistency index.

Visual inspection of co-occurring CLOGs indeed reveals functional relationship. Among significantly correlated pairs are subunits of protein complexes like the cytochrome bd plastoquinol oxidase found in CLOGs 11458 and 11459, further *hypA* (CLOG 10002) and *hypE* (CLOG 12374) as well as *hypE* and *hypF* (CLOG 11744), three subunits of the hydrogenase maturation protein Hyp. Interestingly, while the subunits of the plastoquinol oxidase form an operon structure in all organisms, genes of Hyp can only be found in close genomic proximity in about half (23 of 39) of the species. Another class of correlated CLOGs are genes involved in one metabolic pathway, for example, tocopherol cyclase (EC 5.5.1.24, CLOG 9703) and homogentisate phytyltransferase (EC 2.5.1.115, CLOG 10825). Both CLOGs also co-occur with the 4-hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27, CLOG 10837). Although these three genes are essential for the biosynthesis of Vitamin E, they are not in close genomic proximity on any strain. The examples demonstrate two aspects of co-occurring CLOGs: functional relationships can go beyond pairs and may involve groups of CLOGs; genes of co-occurring CLOGs are not necessarily in close proximity on the genomes. Systematic analysis of co-localization of genes has been successfully applied to identify functional relations (Fouts et al. 2012; Winter et al. 2016), and thus can strengthen the confidence in the validity of co-occurrence. However, our data also shows that co-occurrence provides additional information of functional relationship that can not be observed with co-localization.

Functionally linked metabolic genes most likely concur in metabolic pathways, as they cooperatively work towards the same goal. For a systematic analysis, we used the KEGG database to link all EC numbers to their corresponding metabolic pathways, excluding the very generic *metabolic pathway*. CLOGs were attributed with all pathways of their assigned EC numbers. For all pairs of significantly correlated CLOGs assigned to at least one pathway, we calculated the ratio of shared pathways. As depicted in Figure 4.3A, the ratio of common pathways is below 20% for most CLOG pairs with an SI below 0.5. However, pairs of CLOGs with strong co-occurrence (SI $>$ 0.6) show high similarity in their assigned pathways of up to

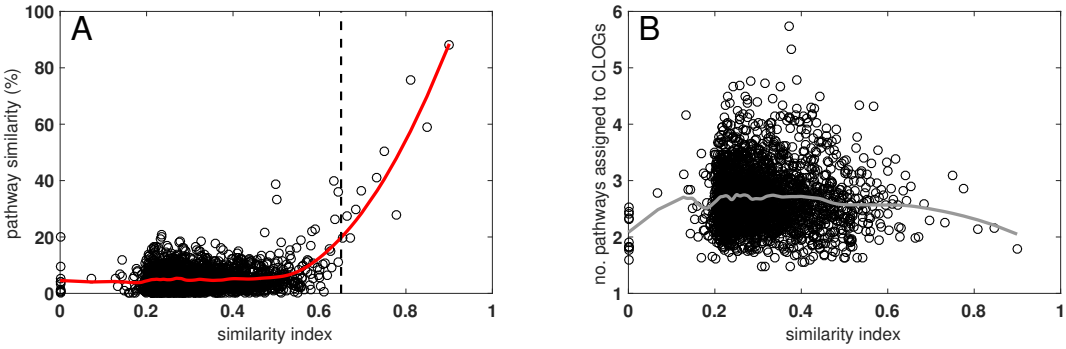


Figure 4.3.: Conformity of metabolic function in co-occurring CLOGs. **A**: The left scatter plot depicts the ratio of shared pathways for all significantly correlated pairs of metabolic CLOGs assigned to a pathway as function of their similarity index. Each circle represents the average of 25 pairs, while the red line indicates a LOESS regression using a span of 25% of the total number of data points. The 0.65-cut-off for similarity used in this thesis is represented by a dashed line. The graph shows that functional relationship, indicated by shared assignment to metabolic pathways, is significantly increased for CLOG pairs with an SI above 0.6. **B**: This effect can not be attributed to chance, as the numbers of assigned pathways for pairs of metabolic CLOGs is effectively equal irrespective of their similarity. Again, each circle represents the average of 25 pairs, while the gray line shows the LOESS regression.

90%, indicating a functional relation of the genes. This effect can not be explained by the number of assigned pathways. CLOGs involved in pairs with high similarity are linked to the same average number of pathways as CLOGs involved in pairs with relatively low SI (Figure 4.3B). Based on the data we set the threshold for similarity at 0.65 and assume that pairs with an SI above that threshold have a high probability for functional linkage.

In addition to co-occurrence, we also studied anti-occurrence, which means pairs of CLOGs associated with mutually excluding subsets of organisms. Or in other words, genes in one CLOG explicitly do not occur in organisms participating in the other CLOG. Compared to co-occurrence, negative correlation is less common. Using the left-tailed Fisher's exact test, again correcting for multiple testing with the method by Benjamini and Yuketiele (Benjamini and Yekutieli 2001) and an accepted FDR of 0.01, we identified 178,408 significantly anti-correlated pairs of CLOGs, which is about one third of the number of correlated pairs. We again calculated the adjusted mutual information, however determining the consistency index is not possible for anti-occurring CLOGs. We note that AMI is non-negative and symmetric with regard to the observation, thus perfectly anti-correlated CLOGs also receive an AMI of one. Among the negative correlated pairs with highest AMI are a dethiobiotin synthase (CLOG 6775) and a hypothetical gene (CLOG 8906), a alpha/beta hydrolase fold (CLOG 4535) and a helicase enzyme (CLOG 6308), as well as a transporter membrane component (CLOG 5269) and a carboxysome associated protein (CLOG 5503). As indicated by the examples, we could not detect systematic functional linkage of negatively correlated CLOGs. Moreover, because we could not

correct for the phylogeny, anti-correlated CLOGs show a dominant association with particular phylogenetic clades. They are typically associated exclusively with α- or β-cyanobacteria. Members of these two clades are distinguished by the molecular structure of the RuBisCO protein and their carboxysomes. α-cyanobacteria utilize a specific 1A form of RuBisCO and distinct α-carboxysomes, which they most likely acquired through horizontal gene transfer from proteobacteria (Badger and Price 2003; Whitehead et al. 2014). Both groups have separated around one billion years ago (Blank and Sánchez-Baracaldo 2010). Anti-correlated CLOGs therefore more likely reflect fundamental differences in the biology of α- and β-cyanobacteria rather than specific mutually excluding biological processes. In the following, we therefore focus only on co-occurrence of CLOGs.

4.3.2. Network analysis of co-occurring CLOGs

Examining pair-wise co-occurrences of CLOGs is not sufficient to fully reveal the underlying structure of functionally related genes. As pointed out above, groups of multiple CLOGs can have high mutual similarity, indicating a biological process involving more than two genes. We identified groups of co-occurring CLOGs, hereinafter denoted as modules, by constructing an undirected network graph. CLOGs were considered as nodes, connected by a weighted link if the co-occurrence between two CLOGs is significant. Weight of the edges $W(i, j)$ is basically the similarity score cut-off at a minimal value of 0.65. Modules of interconnected CLOGs were extracted using the community identification algorithm proposed by Blondel and colleagues (Blondel et al. 2008). This method uses a heuristic approach to maximize the modularity in a given graph and is parameter-free. The results are highly robust with respect to different choices of the cut-off, as most modules are separated components of the network.

In the end, we identified 563 modules encompassing a total of 1,930 CLOGs. Most modules comprise just two (371) or three (93) CLOGs. Only 21 modules consist of more than ten CLOGs. Despite correcting for the phylogenic relation of associated organisms, large modules typically reflect subgroups of cyanobacteria. The largest module comprises 48 CLOGs occurring in most β-cyanobacteria excluding both *Gloeobacter* strains. Yet annotations are often vague and only few CLOGs seem to have possible functional relationships. Similarly, the 41 CLOGs of the second largest module are mostly associated with the marine *Synechococcus* and two *Prochlorococcus* strains, but have mostly ambiguous annotations and limited functional conformity. However, it is worth mentioning that parts of the module, for example CLOGs 20504, 20506, 20509, and 20510, have vague annotations (Uncharacterized secreted or membrane protein), but are co-localized on the genomes of all associated strains. Limited annotation therefore might prevent the detection of possible relationships. In contrast to these examples, smaller modules are typically not associated with particular clades indicating clear functional linkage between CLOGs.

4.3.3. Co-occurrence and co-localization

In addition to co-occurrence, we evaluated the co-localization of CLOGs participating in a joint module. Co-localization is the conservation of genomic proximity of two

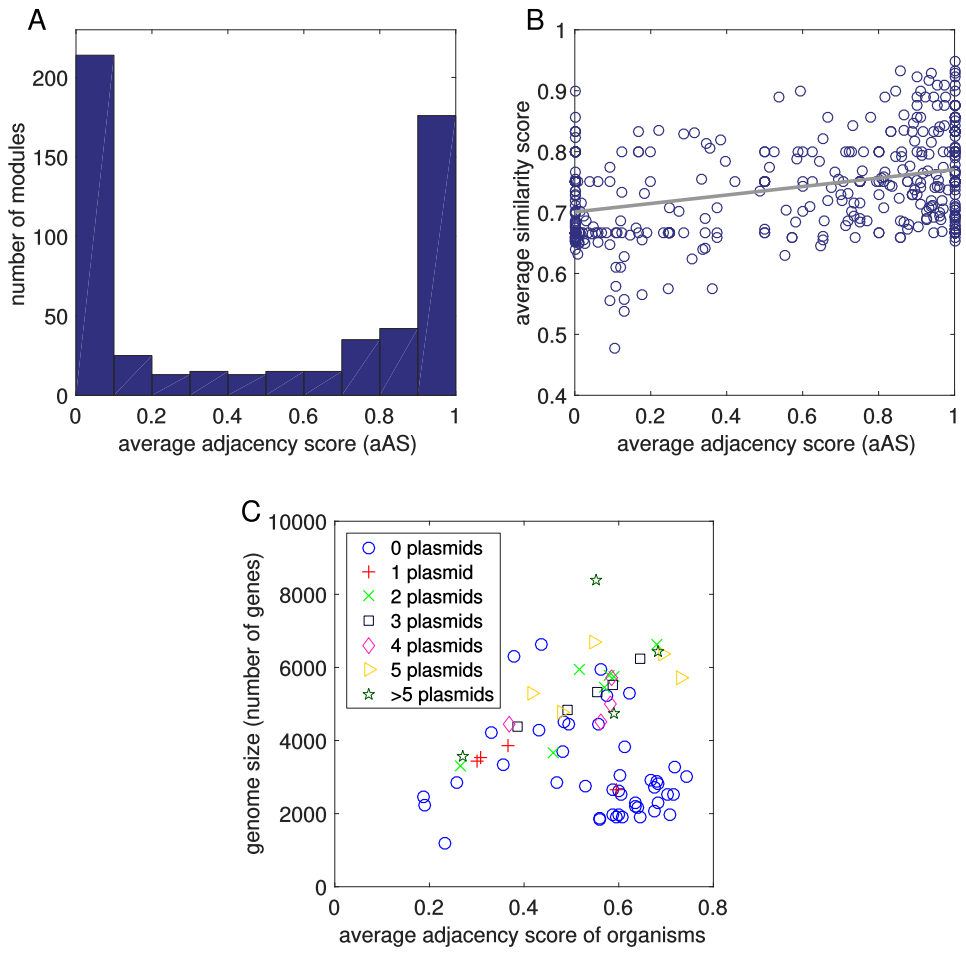


Figure 4.4.: Genomic proximity of co-occurring genes. **A**: The average adjacency score (aAS), representing the proximity of genes grouped in a common module, shows a clear dichotomy. Most modules are either comprising CLOGs, whose genes are adjacent in all genomes, or comprise CLOGs, whose genes are scattered across the genomes. **B**: No clear correlation between the aAS and quality of the co-occurrence, represented by the corrected adjusted mutual information (AMI) on the y-axis, can be observed. The aAS and corrected AMI of each module are represented by an open circle, the linear regression is indicated by a gray line. **C**: The mean adjacency score for all modules of co-occurring genes within a specific strain does neither correlate with the genome size (indicated on the y-axis) nor the number of plasmids (indicated by colored symbols) of the respective strain, although plasmids have been reported to possibly assist in organization of functionally related genes (Jones et al. 1991).

or more genes across phylogenetically distant organisms and has been used to detect functional coherence (Winter et al. 2016). By testing the modules of co-occurring CLOGs for operon-like structures, we can ascertain that co-occurrence does provide new insights into functional linkage beyond co-localized genes.

For each organism and module, we calculated the adjacency score (AS), which is basically the ratio of genes constituting one module being localized within a reasonable distance on the genome of one species. Thus the AS is one, if all corresponding genes within the module are located in close proximity (separated by less than 10 open reading frames) in the respective genome, and goes down to zero, if no two genes are in close proximity to each other. The average adjacency score (aAS) of a module is then given as the average of the AS of all constituent strains. The distribution of the aAS, as depicted in Figure 4.4A, shows a clear bipartition. Most CLOGs consist of genes located in close proximity either in all (aAS˜1) or in none (aAS˜0) of the genomes. Interestingly, the significance of co-occurrence, measured as the average similarity score between CLOGs in one module, only shows negligible correlation with genomic proximity (Figure 4.4B). Genes in CLOGs with low similarity are as likely to be localized in close neighborhood on their genomes as genes of highly similar CLOGs.

The dichotomy of the aAS can be partly explained by the fact that more than half of the modules (371 of 563) are composed of only two CLOGs. Thus, the respective AS can only be either zero or one. Moreover, the genomic proximity is rather conserved across the organisms. For 317 of the 563 modules, we observed identical AS in all associated strains. However, co-localization can also vary drastically between the organisms. For example, module 52 (aAS=0.34) consists of six genes for subunits of the nitrate reductase enzyme (EC 1.7.7.2) and five proteins associated with its assembly. Despite their close functional relationship, the corresponding genes are organized in operon-like structures in only 13 strains (mostly *Synechococcus*), but are spread across the genomes of 21 other strains. Explanation of such differences in the genomic arrangement between strains is not straightforward and might involve the phylogenetic history of the respective genes (Koonin et al. 2001).

To study genomic adjacency on a genome-wide level, the average AS was calculated for all significantly co-occurring genes within each organism. We were unable to observe systematic differences either for small streamlined genomes such as *Prochlorococcus* (García-Fernández et al. 2004) or for strains organizing their genes within multiple plasmids (Figure 4.4C). However, systematic comparison of the strain-specific distributions of AS with the background distribution of the AS across all organisms, revealed significantly lower rates of adjacency for five organisms, including *Cyanobacterium aponinum*, *Cyanobacterium stanieri*, both *Thermosynechococcus* strains, and the model organism *Synechocystis* sp. PCC 6803 (two-sided Kolmogorov-Smirnoff test, p-value < 0.001). Genomes of these cyanobacteria seem to be more fractured than observed for the other strains. In contrast, we were unable to identify organisms with highly structured genomes indicated by adjacency scores significantly larger than the average.

Considering our observations, we conclude that analysis of the genomic neighborhood is not sufficient to determine candidates for functionally related genes. Thorough analysis of co-occurring CLOGs does provide additional insight into the linkage of CLOGs beyond the scope of co-localization.

4.3.4. Modules of co-occurring CLOGs indicate functional relationships

We postulate that modules of co-occurrence are indicative of functional relationship between CLOGs. To test this hypothesis, we will exemplarily discuss 21 of the 563 modules and show that modules in fact provide useful insights into the common functionality of involved genes. Finally, it will be elucidated how thorough analysis of co-occurrence results in novel hypotheses for the biological function of unknown or vaguely annotated genes. Profiles of all CLOGs constituting the discussed modules are depicted in Figure 4.5.

The most apparent instances of functional relationships between CLOGs are subunits of heteromultimeric proteins that co-occur across diverse genomes. For example, **module 249** (2 CLOGs, aAS=0.99) consists of two CLOGs coding for the alpha and beta subunit of a NAD(P)+ transhydrogenase (EC 1.6.1.2) and **module 352** (2 CLOGs, aAS=1) consists of the subunits I and II of cytochrome bd quinol oxidase (EC 1.10.3.14) (Howitt and Vermaas 1998). **Module 67** (5 CLOGs, aAS=0.73) consists of the NiFe-type hydrogenase maturation protein subunits HypA, HypC, HypD, HypE and HybF (Casalot and Rousset 2001). Interestingly, the sixth subunit HypB (CLOG 5882) is not present in this module. This subunit can often be found in multiple copies and is also present in cyanobacteria that do not harbor the other 5 subunits (among others *Leptolyngbya* sp. PCC 7376 and *Synechococcus* sp. CC 9605), suggesting a possible second function in the cell. **Module 82** (4 CLOGs, aAS=0.68) consists of the α, β, and γ subunits of urease (EC 3.5.1.5) as well as the urease accessory protein UreG. Urease catalyzes the hydrolysis of urea into carbon dioxide and ammonia as a source of nitrogen. The protein complex assembly is assisted by the four chaperons UreD, UreE, UreF, and UreG (Carter et al. 2009). For most strains, the accessory proteins UreD, UreE, and UreF are grouped in **module 113** (3 CLOGS, aAS=0.68). The remaining cyanobacteria that possess urease (e.g. *Cyanothece* sp. PCC 7424, *Leptolyngbya* sp. PCC 7376, and *Trichodesmium* IMS101) have a modified UreD and UreF (**module 235**, 2 CLOGS, aAS=0.78) but lag the UreE chaperon. This observation implies that UreD, UreF, and UreG are indispensable for the assembly of urease, while the function of UreE can be supplant by slightly altered UreD and UreF.

A second class of functional relationships is modules, whose constituent CLOGs encode transporters. In cyanobacteria and other gram-negative bacteria, ABC (ATP binding cassette) transporters usually comprise three different molecular components: an ATP-binding/hydrolyzing protein (NBD - nucleotide binding domain), one or two transmembrane proteins (TMD) forming a homo- or heterodimeric structure, and a soluble, secreted substrate-binding protein (BP) (Tomii and Kanehisa 1998). In varying compositions they form a membrane spanning structure that can actively change its conformation to facilitate the transport of various compounds through the membrane (Oldham et al. 2008). **Module 79** (4 CLOGs, aAS=0.87) consists of two transmembrane proteins, one ATP-binding protein, and one soluble substrate-binding protein comprising an ABC transporter with unclear specificity. **Module 102** (3 CLOGS, aAS = 0.87) groups CLOGs for one substrate-binding protein as well as two transmembrane proteins for transport of neutral and charged

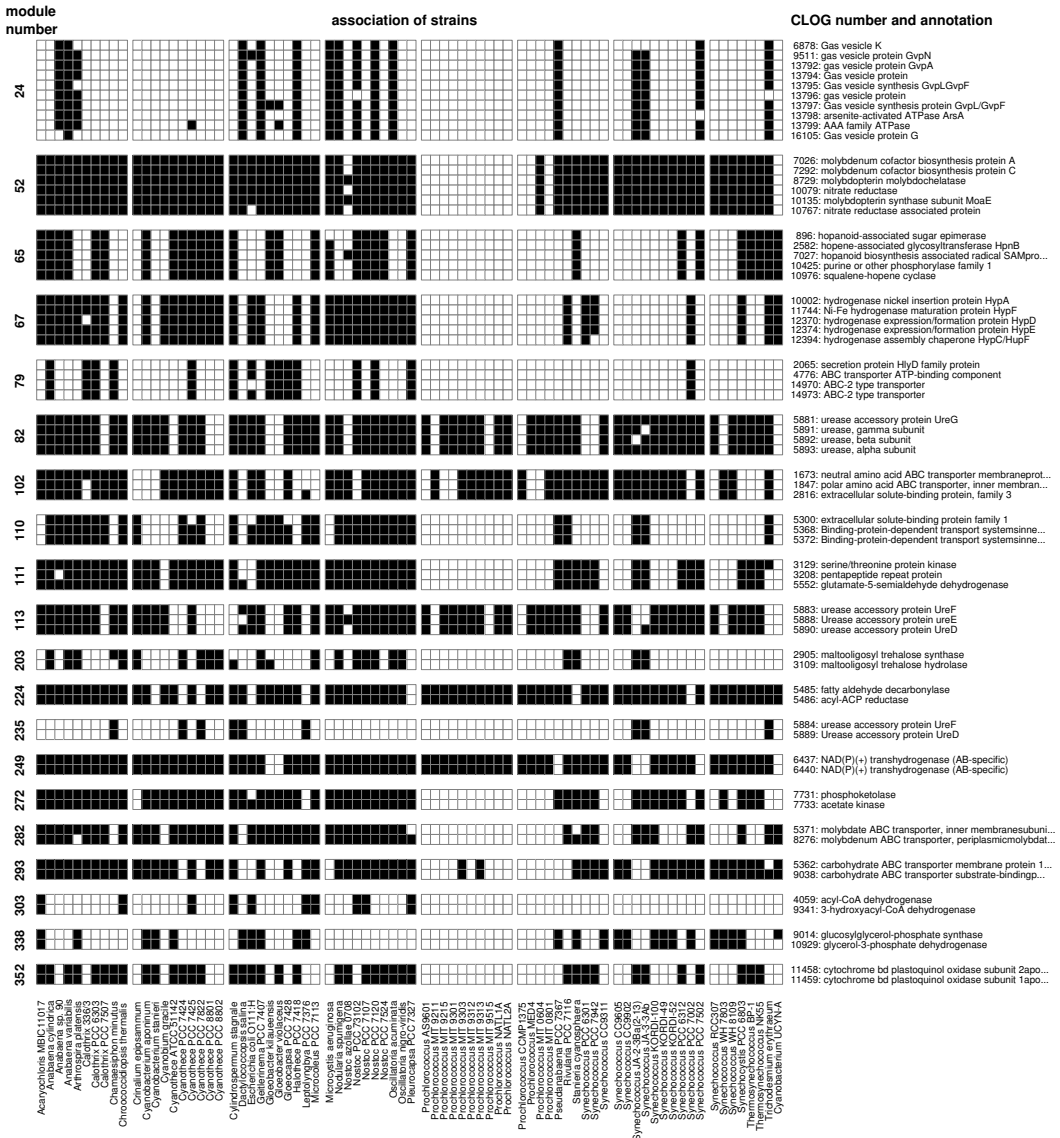


Figure 4.5.: Selected modules and associated strains. For every selected module, the graph shows the profiles of all constituting CLOGs (rows). Black boxes indicate strains (columns) participating in a specific CLOG. Organisms are listed in alphabetical order. The most left column indicates the module number, while the last column shows CLOG number and the respective annotation.

amino acids respectively. These genes are typically found in an operon-like proximity on the genomes together with an ATP-binding protein. **Module 110** (3 CLOGs, aAS=0.62) consists of two transmembrane proteins and one ATP-binding protein forming a putative polyamine transporter; **module 282** (2 CLOGs, aAS=0.65) consists of a molybdate transporter that is assembled from a fused NBD-TMD protein as well as the substrate-binding protein. Interestingly, in 16 of the 47 strains these two genes are not in close proximity on the genome. **Module 293** (2 CLOGs,

aAS=0) consists of the transmembrane and the substrate-binding protein of a carbohydrate transporter. These CLOGs do not form an operon in any strain. We note that ATP-binding NBD proteins are not always modularized with the corresponding transmembrane and substrate-binding proteins. It is known that the ATP-binding proteins of different ABC transporters are highly conserved - up to a degree they can functionally substitute each other (Hekstra and Tommassen 1993; Tomii and Kanehisa 1998). The identity of different NBD-proteins can exceed 60% with a BLAST e-value of 10^{-100} and less. The high degree of sequence similarity therefore results in multiple ATP-binding proteins being clustered together in a few CLOGs (4775 and 4788), compromising the specific patterns of occurrence.

4.3.5. Co-occurrences of CLOGs related to metabolic functions

A third class of functional relationships is modules whose constituent CLOGs encode proteins involved in a common metabolic pathway. For example, **module 272** (aAS=0.4) consists of two CLOGs encoding for a phosphoketolase (EC 4.1.2.22) and an acetate kinase (EC 2.7.2.1), respectively. The phosphoketolase catalyzes the reaction of fructose 6-phosphate to erythrose 4-phosphate and acetyl phosphate, the latter is subsequently converted into acetate by the co-occurring acetate kinase. Therefore, the module reflects a functional association, although both genes are not in close genomic proximity in 32 of 53 strains, including *Synechocystis* sp. PCC 6803. **Module 203** (aAS=0.9) consists of two CLOGs that code for enzymes of the trehalose synthesis pathway, namely maltooligosyl trehalose synthase (EC 5.4.99.15) and maltooligosyl trehalose hydrolase (EC 3.2.1.141) (Higo et al. 2006). **Module 65** (5 CLOGs, aAS=0.31) is associated with the synthesis of hopanoids, pentacyclic compounds that can modify fluidity, permeability, and stability of cell membranes (Kannenberg and Poralla 1999). The module consists of CLOGs, whose genes code for the hopanoid-associated sugar epimerase (HpnA), hopene-associated glycosyltransferase (HpnB), squalene-hopene cyclase (HpnF), hopanoid biosynthesis associated radical SAM protein (HpnH) and a not further specified phosphorylase. All strains participating in this module also harbor at least one copy of the squalene synthase (EC 2.5.1.21), which is also associated with hopanoid synthesis but exists in two variants and is therefore split into the CLOGs 10423 and 10424. **Module 224** (2 CLOGs, aAS=0.91) consists of two CLOGS encoding an aldehyde decarbonylase and an acyl-ACP reductase, respectively. The strict co-occurrence and operon-like structure of both genes has already been described in the context of cyanobacterial alkane biosynthesis (Schirmer et al. 2010; Khara et al. 2013; Klähn et al. 2014). **Module 303** (aAS=0.9) consists of two CLOGs, acyl-CoA dehydrogenase (EC 1.3.8.7) and 3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35), both integral components of the degradation of fatty acids and branched-chain amino acid. Interestingly, in all cyanobacterial strains, genes of these enzymes form an operon-like structure around a gene, which is either part of CLOG 19506 (no clear annotation) or CLOG 9342 (acetyl-CoA acetyltransferase, EC 2.3.1.16). The latter is part of the fatty acids degradation pathway, indicating a similar function of the genes in CLOG 16506. **Module 338** (2 CLOGS, aAS=0.35) consists of glucosylglycerol-phosphate synthase (EC 2.4.1.213) and glycerol-3-phosphate dehydrogenase (EC 1.1.5.3), two

enzymes involved in the synthesis pathway of osmoprotective compound glucosylglycerol, which is produced when the cells are faced with salt stress (Hagemann and Erdmann 1994; Marin et al. 1998). In seven of the 23 strains harboring both CLOGs, the corresponding genes are found in operon-like proximity. Recently, a gene with glycosylglycerol hydrolase activity was identified in *Synechosystis* sp. PCC 6803 (Savakis et al. 2016). The gene is required for re-assimilation of glucosylglycerol, but is not part of the module, as the respective CLOG is annotated in only 13 of the strains considered here, and hence does not strictly co-occur.

4.3.6. Co-occurring CLOGs related to specific cellular functions

The final class of modules combines CLOGs related to specific cellular processes. For example, **module 52** (6 CLOGS, aAS=0.34) consists of CLOGs encoding molybdenum cofactor biosynthesis protein A and C, molybdopterin biosynthesis MoeA and MoeE protein as well as a nitrate reductase and a nitrate reductase associated protein. The co-occurrence can be explained by the co-factor molybdopterin providing molybdenum to the reaction center of the nitrate reductase (Woodard et al. 1990). **Module 24** (8 CLOGs, aAS=0.65) consists of 6 CLOGs related to the assembly of gas vesicles proteins. Gas vesicles allow cyanobacteria a controlled lateral movement in liquid medium. The module also contains CLOGs coding for two ATPases with unknown function that might be involved in vesicle formation or pumping processes. The genes of this module are found in close genomic proximity in 10 of the 16 participating genomes.

4.3.7. Modules provide novel hypotheses for gene function

Of particular interest are modules that combine CLOGs with known function and CLOGs encoding for unknown or putative regulatory proteins. Modules indicating such co-occurrences provide hypotheses about the possible functional role of genes with unknown function and might provide additional insight into regulation of cellular processes. For example **module 111** (3 CLOGs, aAS=0.07) consists of glutamate-5-semialdehyde dehydrogenase (EC 1.2.1.41) involved in the synthesis of essential amino acid L-proline as well as two CLOGs with likely regulatory functions, a pentapeptide repeat protein and a serine/threonine protein kinase. **Module 9** (21 CLOGs, aAS=0.87, data not shown) contains multiple CLOGs associated with the fixation of inorganic nitrogen, as well as five likely regulatory genes. In *Nostoc* sp. PCC 7120 these are asr1405 (hypothetical protein), all1432 (UBA/THIF-type binding protein, probable *hesA*), asl1434 (rop-like domain protein), all2512 (probable transcriptional regulator *patB*), and asr2523 (TPR domain protein). The putative regulatory genes are located almost always in close genomic proximity to the other genes of the module, suggesting a vital role of these genes in the process of nitrogen fixation.

Other modules involve only CLOGs of unknown function and therefore lack a straightforward functional interpretation. In this case, the shared traits of the strains, in which the CLOGs co-occur may provide additional information. For example, **module 3** (aAS=0.24, data not shown) comprises 36 CLOGs mostly annotated

as hypothetical or kinase proteins with 7 signaling-related proteins, two heterocyst differentiation proteins, four membrane transporter related proteins, and two segregation proteins. Genes of module 3 can, with a few exceptions, only be found in filamentous cyanobacterial strains, indicating a role of these genes in filamentous growth. Likewise, **module 4** (aAS=0.11, data not shown) combines 30 CLOGs that are solely associated with filamentous cyanobacteria capable of heterocyst differentiation. The majority of CLOGs in the module lack a specific annotation with only few exceptions, including cytochrome b6f subunit PetM or the heterocyst differentiation protein PatN. It stands to reason that genes in the module are involved in heterocyst formation.

Multiple other modules reveal interesting associations of CLOGs, such as CRISPR-related proteins in module 54, 93, and 97, possible chemotaxis genes in module 62, phosphonate lyase related proteins in module 71, and six transposases in module 50. However, in the scope of this thesis, we could only discuss a small number of modules. The full list of 563 modules is provided in an excel file in Appendix C. To encourage further analysis, we provide the CyanoCLOG SimilarityViewer (SV). This graphical computer program, introduced in Section 4.4, allows the exploration of the co-occurrence neighborhood for any gene of interest. It provides easy access to the corresponding CLOG and facilitates the exploration of co-occurring CLOGs using a variable set of parameters. Summarizing our observations, we conclude that analysis of co-occurrences in a relatively small, yet diverse set of phototrophic organisms is a suitable approach to identify functionally related genes, and therefore to generate novel hypotheses about putative functions of unidentified genes.

4.3.8. Reconstruction of metabolic networks

Reconstruction of genome-scale metabolic networks is a key approach to understand the biosynthetic capabilities of living organisms and therefore is vital for various applications including metabolic engineering. Metabolic reconstructions are typically derived from single genomes, and therefore rely on a correct and complete annotation. By contrast, we used the combined annotations of orthologous genes of 78 organisms, thereby improving the metabolic annotations.

To perform an automated reconstruction of all 78 organisms, we obtained the enzymatic annotation of all genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2008) in combination with a manually curated reconstruction of *Synechosystis* sp. PCC 6803 by Knoop and colleagues (Knoop et al. 2013). CLOGs were annotated with the enzymatic functions of all constituent genes and organisms in return were granted with the metabolic reactions of all associated CLOGs. In addition, a pan-network comprising all reactions was constructed. Every network was complemented with 104 spontaneous reactions. All reactions were tested for chemical balance and fixed to their physiologically feasible directions adopted from the virtual metabolic human database [vmh.uni.lu] (Thiele et al. 2013) and the manual reconstruction of *Synechosystis*.

The pan-network comprised 2,341 distinct reactions and 2,250 metabolic compounds. The number of reactions associated with individual strains ranges from 1,247 for *Prochlorococcus marinus* MIT 9211 to 1,889 for *Nostoc punctiforme* PCC

73102. In cyanobacterium UCYN-A we only identified 943 metabolic reactions, as the organism lives in close endosymbiotic relationship with unicellular algae and lacks the photosystem II as well as various parts of the central metabolism (Thompson et al. 2012). These numbers exceed the number of annotated reactions of other automated metabolic reconstructions, e.g. SEED (Henry et al. 2010) and MetaCyc (Caspi et al. 2014) and are slightly above recently published manually-curated metabolic reconstructions, ranging from 746 distinct reactions for *Arthrospira platensis* NIES-39 (Yoshikawa et al. 2015a) to 851 distinct reactions for *Synechococcus elongatus* PCC 7942 (Triana et al. 2014). A detailed comparison with metabolic reconstructions gathered from MetaCyc and a manually curated network for *Synechocystis* is depicted in Figure 4.6 and reveals three points: roughly half of the metabolic genes have identical annotations; for more than 30%, we identified metabolic functions that were not annotated in the literature models; only few genes (<15%) have assigned enzymatic capabilities in the published models that we could not detect. Most discrepancies can be attributed to annotations with more generic enzyme EC numbers (e.g. generic hexokinase, EC 2.7.1.1, instead of a more specific glucokinase, EC 2.7.1.2) as well as automated and manual gap-filling procedures in the literature models resulting in enzymatic functions of vaguely annotated genes. Other differences are a result of falsely adopting annotations of multi-domain enzymes in phylogenetically distant organisms by MetaCyc, e.g. N5-CAIR synthetase of *Acaryochloris marina* (locus ID: AM1_4761, EC 6.3.4.18) annotated as AIR carboxylase (EC 4.1.1.21) found in higher eukaryotes and malonyl transacylase of *Prochlorococcus marinus* PCC 9215 (locus ID: P9215_01541, EC 2.3.1.39) annotated as fatty-acyl-CoA synthase (EC 2.3.1.86) from yeast. Only a minor fraction of genes was inadequately identified by our method, mainly processes not well covered in the KEGG database and not further investigated in our analyses (e.g. DNA replication, translation, photosynthesis, ATP synthase).

4.3.9. The diversity of cyanobacterial metabolism

The biosynthetic capabilities of the pan-metabolism as well as the metabolic networks of all individual strains was tested by flux balance analysis (FBA). Each network was provided with a minimal medium composed of 19 inorganic nutrients and CO_2 as the sole source of carbon. Because we excluded chemically unbalanced reactions, infeasible accumulation of carbon is prohibited. However, we did not consider energetic restrictions or optimal growth rates of biomass, but simulated the synthesis of each metabolite consecutively. Under these conditions, the cyanobacterial pan-metabolism allows for the production of 881 (of 2,250) metabolites. Figure 4.7 shows the biosynthetic capabilities of individual strains. As expected, genome size (number of genes) and synthesis capabilities are significantly correlated with a Spearman correlation coefficient of roughly 0.79 and a p-value below 0.001 in a permutation test. Strain *Chroococcidiopsis thermalis* PCC 7203 exhibits the largest synthesis potential (683 feasible metabolites, 5,752 genes). In contrast, *Acaryochloris marina* MBIC11017 with the larges genome (8,383 genes) allows for the synthesis of "only" 644 metabolites. Strains with particularly low synthesis potential include, in addition to the minimal strain UCYN-A, *Arthrospira platensis* NIES-39 (564 feasible

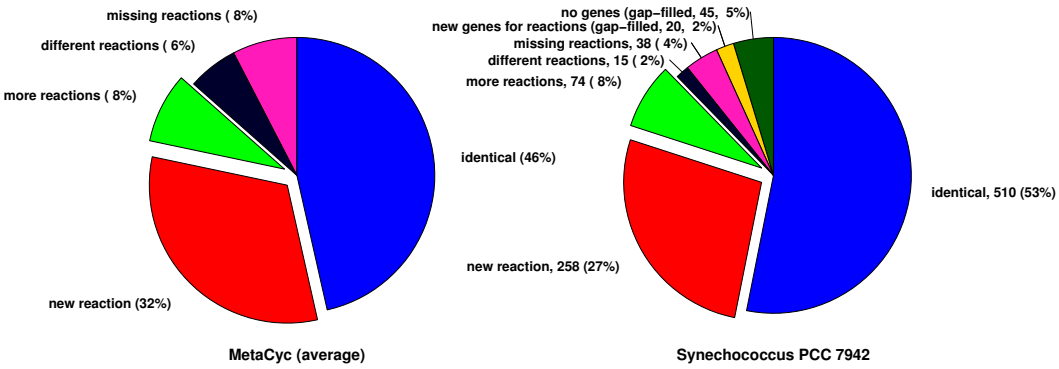


Figure 4.6.: Comparison of reconstructed metabolic networks to published data. The pie charts illustrate the conformity of the metabolic reconstructions to networks provided by MetaCyc (Caspi et al. 2014) and a manually curated network for *Synechococcus elongatus* PCC 7942 by Triana and colleagues (Triana et al. 2014). Metabolic genes were divided into seven categories: identical metabolic annotation (blue), metabolic functions solely identified in our reconstructions (red), additional reactions assigned in our reconstructions (light green), dissimilar annotation in literature and reconstructed networks (black), fewer assigned metabolic functions in our reconstructions (pink), metabolic functions only assigned through manual gap-filling (yellow), metabolic functions identified by means of manual gap-filling but with no identifiable genes (dark green). Genes with amended metabolic annotations in our reconstructions are indicated by exploded slices.

metabolites, 6,630 genes), and *Nostoc azollae* 0708 (488 feasible metabolites, 3,651 genes). The latter is known to live in a perpetual symbiotic relationship with water ferns of the genus *Azollae* (Zheng et al. 2009).

Simulations of the metabolic reconstructions allow us to assess completeness as well as diversity of the cyanobacterial central metabolism. In the scope of this thesis, we consider the synthesis of 46 components related to cyanobacterial growth. These components include all 20 proteinogenic amino acids, RNA and DNA nucleotides, peptidoglycan, glycerol 3-phosphate, several fatty acids, the pigments chlorophyll, β-carotene, γ-carotene, zeaxanthin, and echinenone, as well as the vitamins tocopherol and phylloquinone. Synthesis of all compounds is feasible in the pan-metabolic reconstruction. Biosynthetic capacity of all 78 strain-specific networks for these 46 compounds is depicted in Figure 4.8. Synthesis rates for all 881 metabolites producible by the pan-network are presented in Appendix B. In addition, selected pathways of the compounds discussed in detail are particularized in Figure 4.9.

Most of the metabolic reconstructions possess all relevant enzymes to synthesize proteinogenic amino acids, with the exception of methionine, asparagine, and (iso)leucine. Synthesis of methionine remains enigmatic in several cyanobacteria including *Synechocystis*, although the strain is known to be able to grow in methionine-free media (Gophna et al. 2005; Tanioka et al. 2009). Likewise, key enzymes for the synthesis of asparagine have not been identified for many cyanobacteria (Knoop et al. 2013). The synthesis of asparagine, however, might not be essential for growth

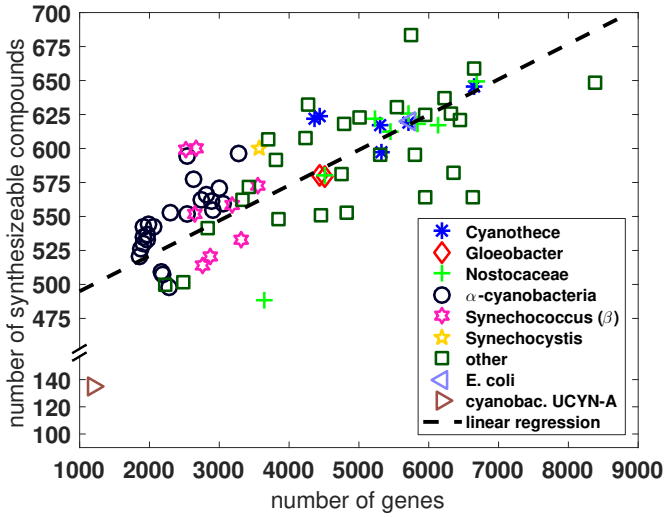


Figure 4.7.: Estimated synthesis capabilities of cyanobacterial strains. Shown is the number of metabolic compounds, for which stoichiometrically balanced synthesis is feasible. Each strain was provided with a minimal set of inorganic nutrients and CO_2 as sole carbon source. Color and form of the marker indicate the most common genera of cyanobacteria, where dark circles denote all α-cyanobacteria including marine *Synechococcus* strains, and pink stars indicate the remaining β-cyanobacterial *Synechococcus* strains. Escherichia coli and cyanobacterium UCYN-A were provided with α-D-Glucose since they lack genes for a fully functional Calvin-Benson-Bassham cycle and cannot assimilate inorganic carbon. The data suggest a strong correlation between genomic size and metabolic capabilities of the strains (Spearman correlation coefficient = 0.79, p-value « 0.001).

because asparaginyl-tRNA can also be synthesized through the transfer of an amido-group to aspartyl-tRNA, using an aspartyl/glutamyl-tRNA amidotransferase (EC 6.3.5.6), which is ubiquitous in cyanobacteria (Curnow et al. 1998). The pathway for the synthesis of leucine and isoleucine via a branched-chain amino acid aminotransferase (EC 2.6.1.42) is not annotated in 15 strain-specific networks but can be (at least from a stoichiometric perspective) replaced by a leucine/isoleucine dehydrogenase (EC 1.4.1.9). However, this enzyme is most likely involved in the degradation of (iso)leucine.

Nucleotides can be synthesized by all strain-specific networks, except UCYN-A, which lacks large parts of the respective pathways. However, no orthologous gene for a ribonucleoside-diphosphate/-triphosphate reductase (EC 1.17.4.1/1.17.4.2) could be identified for six cyanobacteria. In the respective strains, the pathway to synthesize DNA nucleosides is therefore incomplete. Moreover, four organisms lack orthologous genes for thymidylate kinase (EC 2.7.4.9, *Oscillatoria acuminata* and *Synechococcus* sp. PCC 7502) or thymidylate synthase ThyX (EC 2.1.1.148, *Thermosynechococcus*), which are essential for the synthesis of deoxythymidine triphosphate (dTTP). However, genes for ribonucleoside reductase and thymidylate synthase were reported to be hot spots for the insertion of self-splicing introns. Orthol-

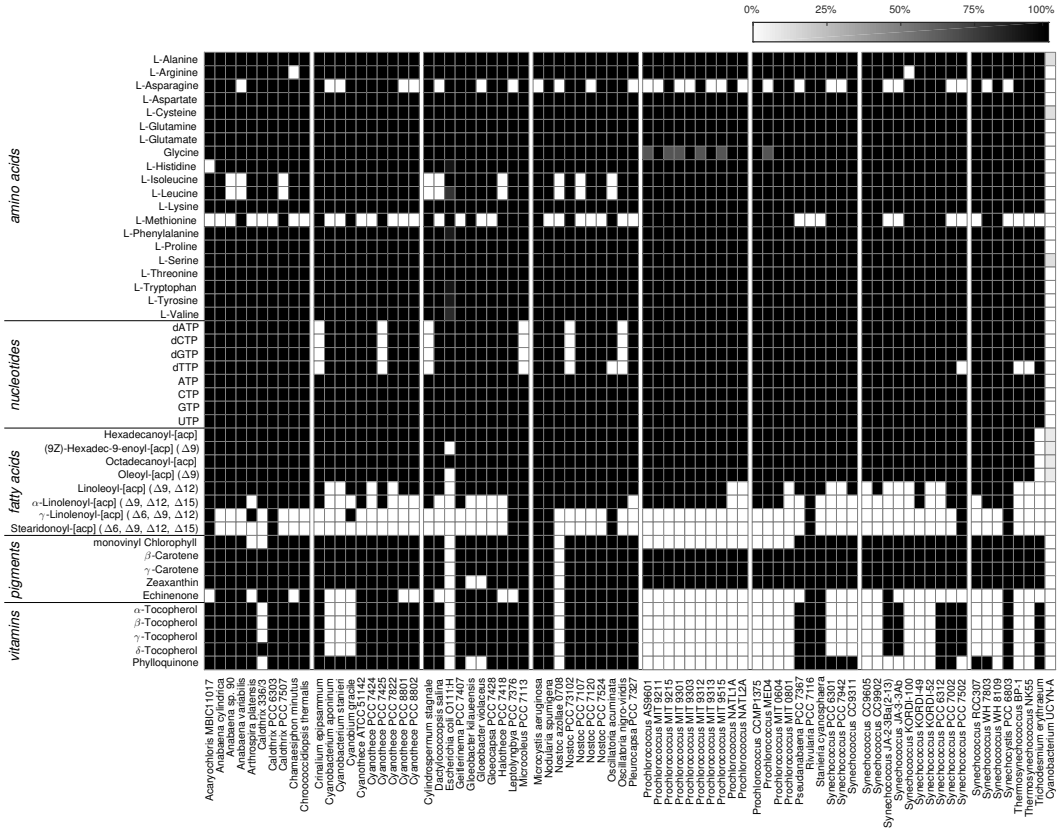


Figure 4.8.: Strain-specific synthesis capacity for 46 growth-related compounds. Shown is the maximal stoichiometric yield of selected growth-related components for strain-specific metabolic reconstructions. The yield is calculated relative to uptake of inorganic carbon. Zero yield (white box) indicates that the respective strain-specific network lacks essential steps for the biosynthesis of this compound.

ogy of these genes is therefore concealed by intervening sequences (Meng et al. 2007; Liu and Yang 2004).

Biosynthesis of fatty acids (Figure 4.9A) is feasible for virtually all strains including E. coli and Cyanobacterium UCYN-A. The fully saturated fatty acids hexadecanoyl and octadecanoyl can therefore be synthesized by all organisms (although in very low quantities by cyanobacterium UCYN-A) with the exception of *Trichodesmium eryhraeum*, which lacks an orthologous gene for the essential S-malonyltransferase (EC 2.3.1.39). In contrast, the presence of desaturases is more diverse, leading to fewer strains capable of synthesizing fatty acids with a high degree of unsaturation as more desaturases are required. The distribution of polyunsaturated fatty acids is in good agreement with a previous study by Chi and colleagues, who compared desaturases in 37 different cyanobacteria (Chi et al. 2008).

The pigment *chlorophyll a* is an essential part of the light harvesting complex of plants and most cyanobacteria. It therefore can be synthesized by almost all strains

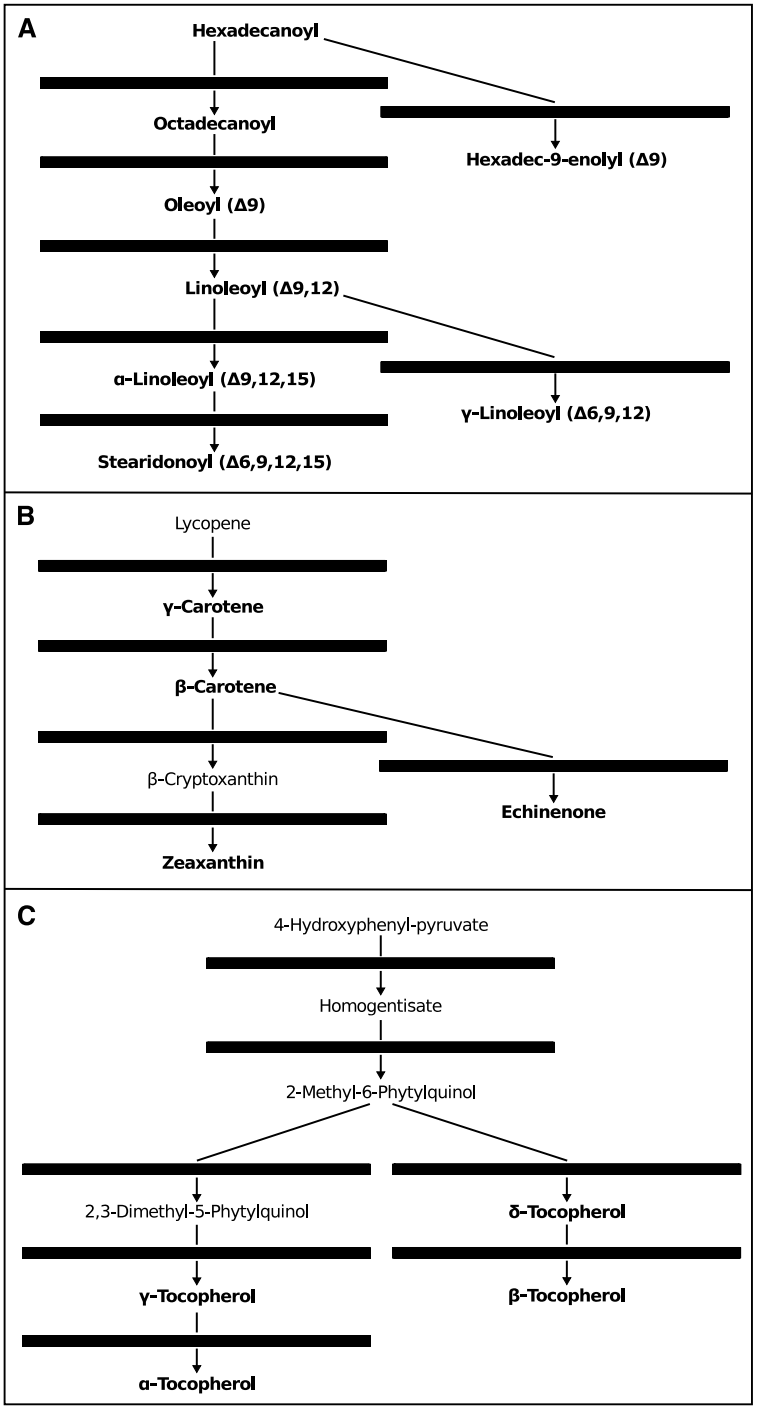


Figure 4.9.: Sketch of selected metabolic pathways. Each panel depicts the consecutive reactions of one metabolic pathway. **A**: Biosynthesis of unsaturated fatty acids, starting from the fully saturated C16 fatty acid hexadecanoly. **B**: Metabolic pathway of several pigments. **C**: Biosynthesis of various tocopherols. Each arrow represents one metabolic reaction. Bars at the arrows show the distribution of the according enzyme in all 78 organisms using the same alphabetic order as in Figure 4.8. Metabolites discussed in the text and specified in Figure 4.8 are emphasized in bold.

with the exception of *Prochlorococcus marinus* as the latter utilize the dissimilar divinyl-chlorophyll and lack orthologs for 8-vinyl-reductase (EC 1.3.1.75) (Ito and Tanaka 2011). The pathway for β-carotene (Figure 4.9B) is present in all cyanobacterial strains except UCYN-A, though *Nostoc azollae* 0708 lacks essential enzymes to synthesize the precursor phytoene. The pathway for zeaxanthin is incomplete in both *Gloeobacter* strains. β-carotene ketolase required for synthesis of Echinenone is not annotated in various strains. The distribution of synthetic capabilities of pigments is well supported by a recent study comparing the carotenoid composition in eleven cyanobacteria (Takaichi and Mochimaru 2007). Takaichi and Mochimaru detected β-carotene in all eleven strains, whereas zeaxanthin could not be observed in *Gloeobacter*. Echinenone was absent or could only be detected at very low levels in strains of *Thermosynechococcus*, *Gloeobacter*, *Prochlorococcus*, and *Synechococcus*.

Three genes comprising the pathway from 4-hydroxyphenylpyruvate to all four variants of tocopherol (Figure 4.9C, also known as vitamin E) are clustered (together with some vaguely annotated regulator and membrane proteins) in module 32. Strains not associated with this module are therefore unable to synthesize tocopherols. Because the genes of this pathway are collectively either absent or present, we have high certainty that the observed synthesis rates are no artifact of flawed annotations. In contrast, the pathway for the synthesis of phylloquinone (vitamin K1) is largely conserved in all strains. However, orthologous genes for enzyme MenH (EC 4.2.99.20) are absent in various organisms including most but not all α-cyanobacteria, *Thermosynechococcus*, and *Gloeobacter* strains. Here, we suspect that incomplete synthesis of phylloquinone is an indication of imperfect understanding of the *menH* genes orthology.

Overall, biosynthesis of the discussed essential components in the strain-specific network reconstructions is in good agreement with published biochemical studies. Despite making no use of any gap-filling method, we detect only few missing orthologous genes in otherwise complete endogenous pathways. Differences in the synthesis of metabolites between strains can most often either be traced back to commonly known flawed understanding of parts of the pathway or be confirmed by previously published biochemical measurements. We therefore assume conclusive predictions regrading the biosynthetic capacities of individual cyanobacterial strains. Interestingly, in contrast to the orthology of the genomes, the pan-metabolism shows no indication of a bimodal structure between α- and β-cyanobacteria. No metabolite can be produced by all α-cyanobacteria but no β-cyanobacteria or vice versa. Almost half of the metabolites synthesizeable by the pan-network (360 of 881) can in fact be synthesized by at least 75 strain-specific networks, accounting for roughly 95%. Less than 11 percent of the compounds (94 of 881) can be synthesized by fewer than five networks ($\sim$5%). All metabolic reconstructions were converted to the universal SBML (Systems Biology Markup Language) format for high interoperability with other software tools and are available from Appendix C. Thus, the networks are available for further research, either aiming at refinement of the networks through automated gap-filling and manual curation, or, because of their predictive power, identification of differences in the metabolic capacities of specific strains, e.g. for bioengineering purposes.

4.4. The SimilarityViewer

As part of this thesis, we discussed more than 20 modules of co-occurring CLOGs. All 58,740 CLOGs and the 563 modules identified with a fixed parameter set (FDR for Fisher's exact test: 0.01; similarity score cut-off: 0.65) are listed in an excel file in Appendix C. For an easy access to this vast set of data and to identify co-occurrence with variable parameters, we created the CyanoCLOG SimilarityViewer (SV). This easy-to-use software tool provides two key functions. For every given gene (defined by the locus ID), it returns the according CLOG and hence the orthologous genes identified in the other organisms. For every CLOG, it facilitates the identification of co-occurring CLOGs with an adjustable set of parameters, thereby allowing the quick exploration of possible functional neighborhood for any gene of interest. The SimilarityViewer is written in MATLAB (by The MatWorks, Inc) and can be accessed from the MATLAB command line. In addition, we compiled standalone versions that can run on computers without MATLAB, requiring an automated installation of the MATLAB Compiler Runtime (MCR). All versions can be downloaded from [http://sourceforge.net/p/similarityviewer/] or are alternatively accessible on Appendix C.

4.4.1. Installation

Installation of the MATLAB command line version is as easy as unpacking the zip file into a folder and running the included *runSV.m* file from the MATLAB command line. This script will load the full data set stored in data.mat into the workspace if necessary and start the SV. Precompiled standalone versions are provided for computers running on Linux, MacOS, and Windows. Executing the SimiliarityViewer installer file appropriate for the operating system will open a dialog window, guiding through the installation process of the SV as well as the required MCR. Installation of the MCR requires an internet connection.

4.4.2. Operating instructions

The SimilarityViewer, depicted in Figure 4.10, provides a simple graphical user interface (UI) combining input and output elements in a single window. The UI is organized in four sections, the central control panel and three surrounding areas. The graph in the left shows the relevant section of the network of co-occurring CLOGs for the current parameter set. The area below shows the assignment of organisms to the selected and co-occurring CLOGs. The table on the right lists the full names of the organisms abbreviated on the bottom plot.

Operation of the SV is straightforward. A single CLOG can be selected by either filling in the CLOG number or a gene locus identifier. In the latter case, the program will automatically identify and select the according CLOG. Four parameters can be specified for customized search of co-occurring CLOGs. First, a drop-down menu allows to select the critical p-value for Fisher's exact test. Options are an FDR of 0.01 using the correction method by Benjamini and Yekutieli (critical uncorrected p-value roughly 1.43×10^{-6}) and uncorrected p-values of 1^{-5}, 0.001, and 0.01. Pairs of CLOGs with larger p-value will be dismissed. Because the similarity score usually

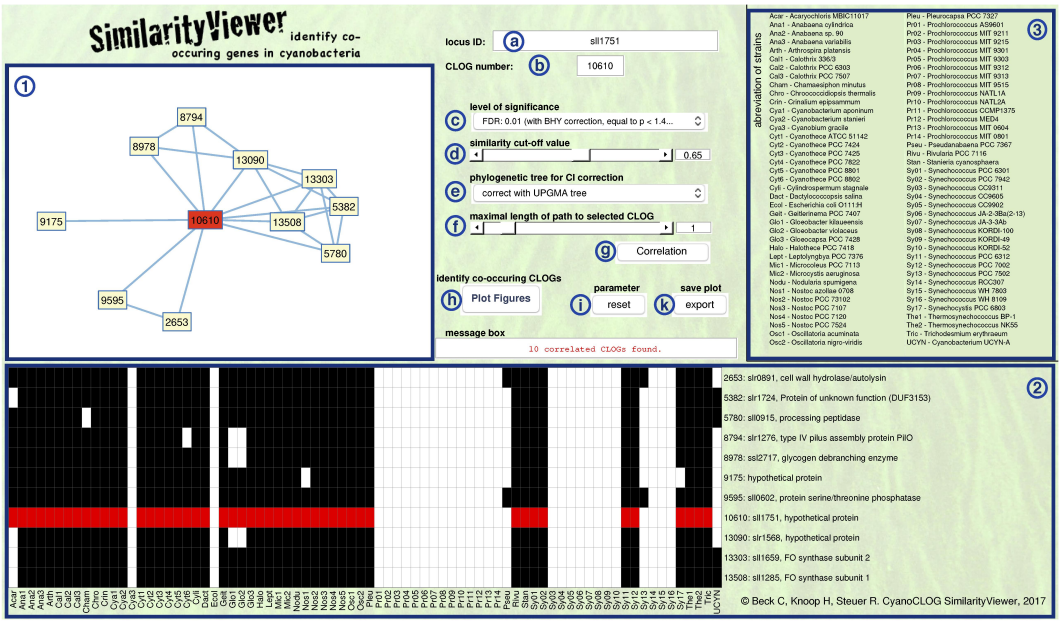


Figure 4.10.: Screenshot showing the SimilarityViewer's graphical user interface. Upon startup, the SimilarityViewer is opened in a single window, which is logically separated into 4 sections. The area on the left (**1**) shows the network of co-occurring CLOGs. Each node represents one CLOG and is connected to other nodes, if the CLOGs are co-occurring according to the selected parameter. The selected CLOG is highlighted in red. The sketch on the bottom (**2**) lists the co-occurring CLOGs (rows) and the associated organisms (columns). On the right, each CLOG is labeled with CLOG number, annotation, and - if possible - the locus IDs of genes associated with the CLOG from the same organism as the selected gene. The selected CLOG again is highlighted in red. Full names of the organisms are given in the table on the right (**3**). The center area features all control elements for an easy exploration of correlated and functionally related CLOGs. A CLOG can be specified either by typing the locus ID of a specific gene into box (**a**), which triggers the search for the corresponding CLOG, whose number is then shown in box (**b**). Alternatively, the CLOG number can be typed into box (**b**) directly. Four parameters can be adjusted using the sliders and drop-down menus, namely the critical value of Fisher's exact test (**c**), cut-off value for the similarity score (**d**), activation/deactivation of the consistency index correcting for phylogenetic imbalance of co-occurring CLOGs (**e**), and the maximal accepted distance in the network of co-occurring CLOGs (**f**). The toggle (**g**) allows to switch between search for correlated and anti-correlated CLOGs. If anti-correlation is activated, calculation of the consistency index is prohibited and the network distance is fixed to one. The main button (**h**) activates the calculation of (anti-)co-occurring CLOGs. Button (**i**) resets all parameters to default values used to identify co-occurring CLOGs in this thesis. Pushing button (**k**) activates the export of the lower sketch (**2**). File name and data type can be adjusted in a pop-up dialog. The message box at the bottom of the control panel shows error messages or useful information on the last operations. For example, when clicking on a square in the bottom sketch, the message box will show the locus ID of the gene associated with the selected CLOG and organism.

is a more stringent parameter, selecting the significance level comes into effect only for low similarity cut-off values. Second, a cut-off for the similarity score can be selected via slider for direct input. CLOGs connected by similarity score below that threshold will be rejected. The cut-off is bound to 0.2 (virtually no similarity) and 1.0 (perfect similarity). Third, correction for phylogenetic imbalance can be adjusted by selecting/deselecting the phylogenetic tree used for the calculation of the consistency index. If deselected, the CI will be set to zero. Fourth, the maximal accepted distance between selected and correlated CLOGs in the network can be adjusted using a slider or direct numeric input. If, for example, set to one, only CLOGs directly linked to the selected CLOG will be accepted. This parameter is bound to zero (shows only the selected CLOG) and ten. In addition to the mentioned parameter, a toggle switch can be activated to search for anti-correlated instead of co-occurring CLOGs. If the computation of anti-co-occurrence is selected, computation of the consistency index is deactivated and the maximal allowed distance is set to one.

Pressing the Plot Figures button will start the computation of co-occurring CLOGs and draw the two plots. The left graph shows the network of co-occurring CLOGs, highlighting the selected CLOG with red color. The plot on the bottom lists all (anti-)co-occurring CLOGs. Associated organisms are indicated by dark boxes. Clicking on one of the boxes will show the locus ID of the gene assigned to the according CLOG and organism in the message box above. Again, the selected CLOG is highlighted in red. Added to each row are the CLOG number and annotation. If a specific gene was used at the beginning to select the original CLOG, the annotation also includes gene locus IDs of co-occurring genes from the same organism.

An export function is included into the SV to save the plot of co-occurring CLOGs in a clean graph for future purposes, to share with colleagues, or for publications. The plot can be saved as pdf or image file. For a clean graphic export, the *export_fig* function provided by Yair Altman [www.mathworks.com/matlabcentral/fileexchange/23629-export-fig] was adapted.

4.5. Discussion

With costs for sequencing of large nucleotide fragments decreasing rapidly, the number of fully sequenced organisms increased accordingly. At the time of this writing, over 8,000 completely assembled prokaryotic genomes are available at the NCBI GenBank database. The vast number of sequences guaranteed high coverage of individual phyla, offering the possibility to investigate the functional relations as well as environmental adaptations at the genomic level of individual species. We compared the genomes of 77 cyanobacterial strains with fully assembled chromosomes gathered from the GenBank database. Utilizing the algorithm introduced in Chapter 3, we identified and clustered all likely orthologous genes (CLOGs). This, in consequence, allowed us to identify and examine the co-occurrence of orthologous genes motivated by the hypothesis that co-occurrence is indicative of a functional relationship between CLOGs.

Cyanobacteria populate the earth for millions of years and form an ancient phylogenetic clade. They show enormous diversity with respect to their preferred environ-

ment (temperature, humidity, concentration of minerals), cell organization (single celled, filamentous, heterocysts forming), and metabolic capability (nitrogen fixation, lipid synthesis, vitamin metabolism). However, almost all cyanobacteria share an autotrophic metabolic lifestyle, using oxygenic photosynthesis as a primary source of energy. We argue that this similar, yet diverse nature of cyanobacterial growth represents an ideal prerequisite for the evaluation of gene co-occurrences. Earlier studies on co-occurrence either focused on a set of very closely related organisms, e.g. 29 E. coli strains (Vieira et al. 2011), or extended the search to a plethora of genomes including all domains of live (Sun et al. 2005; Zhou et al. 2006; Cokus et al. 2007; Kim and Price 2011). In the former case, the set of organisms typically lacks the diversity to simultaneously identify presence and absence of individual genes for particular cellular functions. In the latter case, functional relationships can be obscured by the diversity of metabolism and lifestyle characteristically for a vast number of unrelated prokaryotic and eukaryotic organisms. As indicated by previous studies, adding more genomes does not necessarily benefit the quality of co-occurrence (Snitkin et al. 2006; Jothi et al. 2007; Škunca and Dessimoz 2015). We therefore are confident that the selected 78 prokaryotic organisms represent an ideal set for the identification of co-occurrence with respect to putative functional relationships in cyanobacteria.

To investigate co-occurrences of CLOGs, we developed a customized approach based on a network perspective that allows us to identify modules of co-occurring CLOGs. Our results demonstrate that modules of co-occurring genes are indeed often indicative of functional relationships. Straightforward examples include subunits of heteromultimeric enzymes, often organized in known operon-like structures. Modules of co-occurring genes, however, can also be associated with the catalysis of sequential steps in a metabolic pathway or other specific cellular functions. Examples discussed here were - among several others - the biosynthesis of molybdopterin as well as the assembly of gas vesicle proteins. In addition we showed that functionally associated genes don't have to be organized in close genomic proximity, thus systematic analysis of co-occurrence does supplement the identification of functionally related genes purely based on genomic co-localization (Winter et al. 2016).

As revealed by the discussed examples, co-occurrence results in high confidence of a shared biological function, which we expect holds true for other modules as well. Of particular interest are modules that consist of CLOGs with known specific function as well as CLOGs with ambiguous or unknown annotations, as such modules provide novel hypotheses for putative gene functions. In order to facilitate further analysis of functional neighborhood based on the method presented, we developed a computational toolbox. The SimilarityViewer allows the identification and exploration of co-occurrence for every gene and with varying parameters, beyond the examples discussed in the scope of this thesis. It is available for MATLAB as well as a stand-alone application for Mac, Linux, and Windows at [http://sourceforge.net/p/similarityviewer/] and in Appendix C.

Utilizing revised genomic annotation courtesy of extensive analysis of gene orthology and incorporation of a manually curated metabolic network, we comprehensively reconstructed the metabolism of 78 mostly cyanobacterial strains. Simulations of the networks using flux balance analysis indicated a high integrity of the metabolism.

Providing minimal medium and CO_2 as sole source of carbon, most networks were capable of synthesizing the majority of compounds vital for photosynthetic growth. Discussing the biosynthesis of 46 growth-related metabolites in detail, we showed that differences in synthetic capacities could either be confirmed by published biochemical studies or attributed to a general lack of knowledge of the genes involved in the respective pathways. We assume similar accuracy in predicting the biosynthetic capacities for the other metabolites. For flawless integration into other software tools, all networks were exported to annotated SBML files and are available in Appendix C. We have high confidence that the networks will become a valuable tool for further comparison of biosynthetic capabilities and metabolic pathways in cyanobacteria. Furthermore, they provide an excellent basis for refined manually curated reconstructions of the cyanobacterial metabolism.

Chapter 5.

Summary

In this thesis, we tackled two major questions concerning the lifestyle of phototrophic cyanobacteria: First, how do cyanobacteria organize their intracellular processes in response to the rhythmic changes between day and night. Second, what are the commonalities as well as the diversities between the metabolism of various cyanobacterial strains.

Temporal organization of intracellular processes
To address the first question, we conducted a comprehensive microarray time-series study of the model organism *Synechocystis* sp. PCC 6803, and quantified rhythmic gene expression in changing light-dark conditions as well as continuous light and continuous darkness. Although we were unable to identify any persisting oscillation in continuous conditions, we observed a rigorous schedule of gene expression in day-night conditions. Consistent with the photoautotrophic lifestyle, processes involving photosynthesis, carbon-fixation, or translation are upregulated during the night, while proteins involved in respiration and maintenance are upregulated at the end of the day to be available during the night. We therefore exclude the presence of circadian gene regulation in *Synechocystis* in the slow growing conditions applied in this thesis, although later studies described circadian rhythms in other conditions (van Alphen and Hellingwerf 2015). *Synechocystis* possesses multiple copies of the core circadian clock genes, including a *kaiA-kaiB-kaiC* cluster highly similar to the genes responsible for circadian rhythms in *Synechococcus elongatus* PCC 7942, as well as two more copies each of *kaiB* and *kaiC*. However, upon close inspection we found that most *kai* genes are directly regulated by *cis*-encoded regulatory asRNA. We therefore hypothesize a facultative clock mechanism that can switch between self-sustained circadian oscillations and hourglass-like behavior. The latter can also be observed in and seems to be sufficient for the cyanobacterium *Prochlorococcus*, which possesses no homolog for *kaiA* (Axmann et al. 2009).

In a recent study, we analyzed the occurrence of clock-related genes in organisms from all domains of life (Schmelling et al. 2017). Multiple copies of *kaiB* and *kaiC* can not only be found in *Synechocystis*, but in various other cyanobacterial strains as well. We also observed single or even multiple homologues of these two genes in other bacteria and archaea. In contrast, *kaiA*, which is essential for sustained circadian oscillations of the *kai*-clock, is exclusively present in cyanobacteria. The ability to temporally organize gene expression with an hourglass-like clock mechanism seems to be beneficial for various prokaryotic organisms. However, a true circadian clock is likely exclusive for cyanobacteria and more advanced eukaryotes. In addition, benefits of a circadian clock in cyanobacteria are relatively slim and the induced

regulation is by no means universal across multiple strains. Interrupting the circadian clock results in similar growth rates, within the error of the measurement. Only in direct competition, strains with a clock adjusted to the external light conditions are advantaged and outcompete other strains (Woelfle and Johnson 2009). In our recent study, we also compared diurnal and circadian rhythms reported for various cyanobacterial strains. However, with the exception of a few dawn peaking genes - mainly related to translation or photosynthesis - no homologous genes showed similar expression patterns across all studies (Schmelling et al. 2017). Furthermore, some studies show striking differences between transcriptome and proteome data. Often, protein levels of genes with an oscillatory expression profile are not in phase with the transcript levels but rather constant over time or even antiphasic (Welkie et al. 2014; Guerreiro et al. 2014). These findings imply an unknown post-transcriptional level of rhythmic control that might be fully independent from the circadian clock or controlled by the clock through a yet not understood mechanism. Circadian expression might only give a rough timing pattern of the transcript levels, while final translation is than controlled by other, more refined processes.

We therefore conclude that the multiple copies of *kaiB* and *kaiC* and the strong regulation of *kai* genes through asRNA are indicative of a facultative circadian clock mechanism, to avoid costs associated with maintenance of a sustained oscillator in non-optimal growth conditions. Conditionality of the clock at low temperatures has already been reported for *Synechococcus elongatus* PCC 7942 (Xu et al. 2013a). However, to ascertain conditionality and to identify conditions, in which activation of a self-sustained clock is beneficial, performing further time-series experiments at various growth conditions is necessary.

In addition to the tight regulation of clock genes and the resulting diurnally scheduled gene expression, we observed strong oscillation in the amount of long ribosomal RNAs. 16S and 23S rRNAs accumulated throughout the night and degraded rapidly after transfer to light. This behavior can not be observed in constant light and is, thus, likely driven by the light cycle rather than a circadian mechanism. Although moderate accumulation of RNA has been described before, common conception is that the amount of rRNA is proportional to the growth rate (Lepp and Schmidt 1998; Binder and Liu 1998). Yet in our experiments, growth was negligible during the night. Because the amount of short 5S rRNA and tRNAs is constant over time, we can most certainly exclude global and technical reasons for the fluctuations, such as general accumulation of ribosomes and lower efficiency of the RNA purification in samples from dark phases. We therefore hypothesized that long rRNAs might act as a storage compound for the pentose sugar ribose, which can be quickly cleaved from the nucleotides and converted to ribulose 1,5-bisphosphate, substrate of the CO_2-fixing enzyme RuBisCO. This would prevent the energy-intensive Calvin-Benson-Bassham cycle from running during the night, where photosynthesis is not possible, but allow for a quick restart of the growth-determining carbon fixation process right at the beginning of the day. Therefore, cells would not need to activate enzymes related to the Calvin-Benson-Bassham cycle before the transition to light, which presumably reduces the need for a true circadian clock. Nightly accumulation of rRNA might be a requisite of a reduced diurnal clock system. However, to fully understand the biological processes underlying rapid rRNA degradation, conducting further time-series

experiments measuring enzyme activity and metabolic fluxes during the transition from dark to light is necessary.

Metabolic diversity of cyanobacteria

To address the second question about genomic and metabolic diversity in cyanobacteria, we established a new method for a genome-wide genetic comparison of multiple cyanobacterial strains by clustering likely orthologous genes. Each CLOG comprises functionally similar genes present in one or multiple organisms. Upon close inspection, the distribution of CLOG size exhibits a typical pattern. Most genes occur in only one or few organisms, whereas only few genes are shared by a large subgroup of the organisms. However, about 620 to 660 core genes are shared by all photoheterotrophic cyanobacteria, which leads us to believe that around 600 genes are common in most cyanobacterial strains. We found no evidence of a closed cyanobacterial pan-genome, thus *de novo* sequencing of cyanobacterial genomes will inevitably lead to the discovery of novel genes.

Core CLOGs represent an indispensable set of genes. Thorough analysis of the functional annotations revealed that core CLOGs are predominantly involved in housekeeping functions, such as various metabolic processes and translation. CLOGs shared by only few organisms, in contrast, are mostly related to processes involved in adaptation and individual cellular defense. Accordingly, investigating core metabolic pathways revealed rather strong conservation of enzymes involved in the pentose phosphate pathway, the Calvin-Benson-Bassham cycle, glycolysis, as well as the catabolism and anabolism of glycogen. Other central pathways, however, are less conserved across the strains and rather tailored to the organisms' demands. For example, the TCA cycle is highly reduced in α-cyanobacteria, allowing no cyclic flux. Yet, parts of the cycle, obligatory for the biosynthesis of amino acids, are present in these organisms as well, while anaplerotic and cataplerotic reactions were adjusted accordingly. Biosynthesis of cyanophycin and poly-β-hydroxy-butyrate can only be observed in a subset of the organisms, thus genes involved in the synthesis of the particular storage compound are concertedly present in only these organisms. As indicated by these examples, co-occurrence of genes can be attributed to a common function.

Going one step further, we thoroughly investigated co-occurrence of genes in a wide set of cyanobacterial strains. By developing a novel network-based algorithm, we grouped pairs of co-occurring genes into modules, which indeed were indicative of functional relationship among the genes. Modules identified in this thesis can be divided into four categories. First, modules comprising multiple subunits of heteromultimeric enzymes such as the hydrogenase enzyme and ABC transports. Second, modules of proteins collectively involved in the formation of complex intracellular structures, e.g. assembly of gas vesicles. Third, modules of genes associated with the catalysis of sequential steps in a metabolic pathway such as biosynthesis of molybdopterin. Fourth, modules whose genes have no obvious relationship but co-occur in organisms with specific traits, for example the ability to form heterocyst cells. High levels of functional relationship within modules of co-occurring genes is also expected for modules not discussed in this thesis. This is particularly interesting for modules consisting of genes with known functions and with ambiguous annotations,

as such can provide hypotheses for putative gene functions. To facilitate further analysis of functional neighborhoods we developed the SimilarityViewer, allowing easy exploration of co-occurrence for every gene.

By utilizing revised genomic annotations and incorporating annotations from a comprehensive manually curated metabolic reconstruction, we were able to automatically reconstruct the metabolism of 77 cyanobacterial strains. For subsequent simulation of the networks using flux balance analysis, chemically unbalanced reactions were rejected and only minimal medium with CO_2 as sole source of carbon was provided. All but two networks were capable of synthesizing most compounds essential for photosynthetic growth and differences in synthesis rates of 46 exemplarily discussed metabolites could either be confirmed by published biochemical studies or attributed to a general lack of knowledge of the genes involved in the respective pathways. Similar accuracy can be assumed for the calculation of the synthesis of other metabolites, shown in Appendix B, thus making our reconstructions valuable for further investigations of the biosynthetic capabilities of cyanobacteria.

With constant increase of the world's population, the demand for agricultural products and fossil fuels rises accordingly. At the same time, the area of arable land is constant and the remaining amounts of fossil fuels are decreasing, while their use becomes less and less acceptable due to high pollution and their contribution to the global climate change. Because of their unique properties, cyanobacteria emerged as a credible alternative for sustainable production of biofuel and various commodities such as polymers, pharmaceuticals, and fodder. Pushed by the urge to drastically reduce greenhouse gas emissions, commercial utilization of cyanobacteria is investigated all over the world. New developments and applications are published frequently in this expanding field of science (e.g. Woo and Lee 2017; Singh et al. 2017). With our work we contributed to the effort of thoroughly understanding cyanobacteria and their unique metabolisms. Comprehensive knowledge about groups of co-occurring and thereby likely functionally related genes is provided by the SimilarityViewer and will be beneficial for the discovery of novel gene functions and economically advantageous metabolic pathways. Our metabolic reconstructions, while no perfect representation of the strains' metabolism, allow the quick screening of biosynthetic capabilities of the individual cyanobacteria strain and help in identifying common metabolic pathways. In addition, they provide a sound foundation for detailed metabolic analyses and manual curation of cyanobacterial metabolic networks.

Contributions

The microarray time-series experiments would not have been possible without the help of various people. Sampling of the RNA was accomplished in collaboration with many colleagues including Ilka Axmann, Beate Heilmann, Stefanie Hertel, Adrian Kölsch, Anne Rediger, Michael Tillich, Anika Wiegard, and Reimo Zoschke. Purification of RNA and preparation for the Agilent Bioanalyzer experiments were mainly conducted by Anne Rediger with assistance by Stefanie Hertel and Adrian Kölsch. RNA labeling and microarray hybridization was carried out by Gudrun Krüger from the Department of Genetics & Experimental Bioinformatics at the University of Freiburg. Robert Lehmann helped with the clustering of the expression profiles. Analysis of metabolic diversity of cyanobacteria was assisted by Henning Knoop with his extensive knowledge of the metabolism in *Synechocystis* sp. PCC 6803.

Bibliography

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS Comput Biol*, 9(3):e1002980.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinforma*, 22(13):1600–7.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

Anderson, S. L. and McIntosh, L. (1991). Light-activated heterotrophic growth of the cyanobacterium synechocystis sp. strain pcc 6803: a blue-light-requiring process. *J Bacteriol*, 173(9):2761–7.

Angermayr, S. A., Hellingwerf, K. J., Lindblad, P., and de Mattos, M. J. T. (2009). Energy biotechnology with cyanobacteria. *Curr Opin Biotechnol*, 20(3):257–63.

Angermayr, S. A., Rovira, A. G., and Hellingwerf, K. J. (2015). Metabolic engineering of cyanobacteria for the synthesis of commodity products. *Trends Biotechnol*, 33(6):352–61.

Aoki, S., Kondo, T., and Ishiura, M. (1995). Circadian expression of the dnak gene in the cyanobacterium synechocystis sp. strain pcc 6803. *J Bacteriol*, 177(19):5606–11.

Aoki, S., Kondo, T., Wada, H., and Ishiura, M. (1997). Circadian rhythm of the cyanobacterium synechocystis sp. strain pcc 6803 in the dark. *J Bacteriol*, 179(18):5751–5.

Aoki, S. and Onai, K. (2009). *Circadian Clocks of Synechocystis sp. Strain PCC 6803, Thermosynechococcus elongatus, Prochlorococcus spp., Trichodesmium spp. and Other Species*, pages 259–282. Springer Berlin Heidelberg, Berlin, Heidelberg.

Apic, G., Gough, J., and Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 310(2):311–25.

Asada, Y., Miyake, M., Miyake, J., Kurane, R., and Tokiwa, Y. (1999). Photosynthetic accumulation of poly-(hydroxybutyrate) by cyanobacteria – the metabolism and potential for co2 recycling. *Int J Biol Macromol*, 25(1-3):37–42.

Axmann, I. M., Dühring, U., Seeliger, L., Arnold, A., Vanselow, J. T., Kramer, A., and Wilde, A. (2009). Biochemical evidence for a timing mechanism in prochlorococcus. *J Bacteriol*, 191(17):5342–7.

Axmann, I. M., Hertel, S., Wiegard, A., Dörrich, A. K., and Wilde, A. (2014). Diversity of kaic-based timing systems in marine cyanobacteria. *Mar Genom*, 14:3–16.

Badger, M. R. and Price, G. D. (1994). The role of carbonic anhydrase in photosynthesis. *Annu Rev Plant Biol*, 45(1):369–392.

Badger, M. R. and Price, G. D. (2003). Co2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot*, 54(383):609–22.

Balaji, S., Gopi, K., and Muthuvelan, B. (2013). A review on production of poly β hydroxybutyrates from cyanobacteria for the production of bio plastics. *Algal Res*, 2(3):278–85.

Bandyopadhyay, A., Elvitigala, T., Welsh, E., Stöckel, J., Liberton, M., Min, H., Sherman, L. A., and Pakrasi, H. B. (2011). Novel metabolic attributes of the genus cyanothece, comprising a group of unicellular nitrogen-fixing cyanothece. *MBio*, 2(5).

Barker, D., Meade, A., and Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinforma*, 23(1):14–20.

Baroukh, C., Muñoz-Tamayo, R., Steyer, J.-P., and Bernard, O. (2015). A state of the art of metabolic networks of unicellular microalgae and cyanobacteria for biofuel production. *Metab Eng*, 30:49–60.

Bassham, J. A., Benson, A. A., Kay, L. D., Harris, A. Z., Wilson, A. T., and Calvin, M. (1954). The path of carbon in photosynthesis. xxi. the cyclic regeneration of carbon dioxide acceptor1. *J Am Chem Soc*, 76(7):1760–1770.

Beck, C., Hertel, S., Rediger, A., Lehmann, R., Wiegard, A., Kölsch, A., Heilmann, B., Georg, J., Hess, W. R., and Axmann, I. M. (2014). Daily expression pattern of protein-encoding genes and small noncoding rnas in synechocystis sp. strain pcc 6803. *Appl Environ Microbiol*, 80(17):5195–206.

Beck, C., Knoop, H., Axmann, I. M., and Steuer, R. (2012). The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genom*, 13(1):1–17.

Beck, C., Knoop, H., and Steuer, R. (2018). Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between groups of ortholog genes. *PLoS Genet*, 14(3):e1007239.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals Stat*, 29(4):1165–1188.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). Genbank. *Nucleic Acids Res*, 36(Database issue):D25–30.

Berman-Frank, I., Lundgren, P., and Falkowski, P. (2003). Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Res Microbiol*, 154(3):157–164.

Binder and Liu (1998). Growth rate regulation of rrna content of a marine synechococcus (cyanobacterium) strain. *Appl Environ Microbiol*, 64(9):3346–51.

Blank, C. E. and Sánchez-Baracaldo, P. (2010). Timing of morphological and ecological innovations in the cyanobacteria – a key to understanding the rise in atmospheric oxygen. *Geobiol*, 8(1):1–23.

Blankenship, R. E. (2001). Molecular evidence for the evolution of photosynthesis. *Trends Plant Sci*, 6(1):4–6.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J Stat Mech-Theory E*, 10:P10008.

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet*, 15(2):107–20.

Bremer, H. and Dennis, P. P. (2008). Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3(1).

Bricker, T. M., Zhang, S., Laborde, S. M., Mayer, 3rd, P. R., Frankel, L. K., and Moroney, J. V. (2004). The malic enzyme is required for optimal photoautotrophic growth of synechocystis sp. strain pcc 6803 under continuous light but not under a diurnal light regimen. *J Bacteriol*, 186(23):8144–8.

Buchanan, B. B., Gruissem, W., and Jones, R. L. (2015). *Biochemistry and Molecular Biology of Plants*. Wiley-Blackwell, 2nd edition.

Carmo-Silva, E., Scales, J. C., Madgwick, P. J., and Parry, M. A. J. (2015). Optimizing rubisco and its regulation for greater resource use efficiency. *Plant Cell Environ*, 38(9):1817–32.

Carpenter, E. J. (2002). Marine cyanobacterial symbioses. *Biol Environ Proc Royal Ir Acad*, 102B(1):15–18.

Carter, E. L., Flugga, N., Boer, J. L., Mulrooney, S. B., and Hausinger, R. P. (2009). Interplay of metal ions and urease. *Met*, 1(3):207–21.

Casalot, L. and Rousset, M. (2001). Maturation of the [nife] hydrogenases. *Trends Microbiol*, 9(5):228–37.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 42(Database issue):D459–71.

Cavalier-Smith, T. (2000). Membrane heredity and early chloroplast evolution. *Trends Plant Sci*, 5(4):174–182.

Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol*, 52(Pt 1):7–76.

Chen, X., Schreiber, K., Appel, J., Makowka, A., Fähnrich, B., Roettger, M., Hajirezaei, M. R., Sönnichsen, F. D., Schönheit, P., Martin, W. F., and Gutekunst, K. (2016). The entner-doudoroff pathway is an overlooked glycolytic route in cyanobacteria and plants. *Proc Natl Acad Sci U S A*, 113(19):5441–6.

Chi, X., Yang, Q., Zhao, F., Qin, S., Yang, Y., Shen, J., and Lin, H. (2008). Comparative analysis of fatty acid desaturases in cyanobacterial genomes. *Comp Funct Genom*, page 284508.

Chisholm, S. W., Frankel, S. L., Goericke, R., Olson, R. J., Palenik, B., Waterbury, J. B., West-Johnsrud, L., and Zettler, E. R. (1992). Prochlorococcus marinus nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. *Arch Microbiol*, 157(3):297–300.

Chisti, Y. (2007). Biodiesel from microalgae. *Biotechnol Adv*, 25(3):294–306.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression-analysis by local fitting. *J Am Stat Assoc*, 83(403):596–610.

Coates, R. C., Podell, S., Korobeynikov, A., Lapidus, A., Pevzner, P., Sherman, D. H., Allen, E. E., Gerwick, L., and Gerwick, W. H. (2014). Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PLoS ONE*, 9(1):1–12.

Cogne, G., Gros, J.-B., and Dussap, C.-G. (2003). Identification of a metabolic network structure representative of arthrospira (spirulina) platensis metabolism. *Biotechnol Bioeng*, 84(6):667–76.

Cogne, G., Rügen, M., Bockmayr, A., Titica, M., Dussap, C.-G., Cornet, J.-F., and Legrand, J. (2011). A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: Application to the green algae chlamydomonas reinhardtii. *Biotechnol Prog*, 27(3):631–40.

Cokus, S., Mizutani, S., and Pellegrini, M. (2007). An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinforma*, 8 Suppl 4:S7.

Curnow, A. W., Tumbula, D. L., Pelaschier, J. T., Min, B., and Soll, D. (1998). Glutamyl-trna(gln) amidotransferase in deinococcus radiodurans may be confined to asparagine biosynthesis. *Proc Natl Acad Sci U S A*, 95(22):12838–43.

Dal'Molin, C. G. d. O., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010). C4gem, a genome-scale metabolic model to study c4 plant metabolism. *Plant Physiol*, 154(4):1871–85.

Das, P., Thaher, M. I., Hakim, M. A. Q. M. A., and Al-Jabri, H. M. S. (2015). Sustainable production of toxin free marine microalgae biomass as fish feed in large scale open system in the qatari desert. *Bioresour Technol*, 192:97–104.

Date, S. V. and Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol*, 21(9):1055–62.

Dismukes, G. C., Carrieri, D., Bennette, N., Ananyev, G. M., and Posewitz, M. C. (2008). Aquatic phototrophs: efficient alternatives to land-based crops for biofuels. *Curr Opin Biotechnol*, 19(3):235–40.

Dittmann, E., Gugger, M., Sivonen, K., and Fewer, D. P. (2015). Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. *Trends Microbiol*, 23(10):642–52.

Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N. J., Angiuoli, S. V., Oggioni, M., Dunning Hotopp, J. C., Hu, F. Z., Riley, D. R., Covacci, A., Mitchell, T. J., Bentley, S. D., Kilian, M., Ehrlich, G. D., Rappuoli, R., Moxon, E. R., and Masignani, V. (2010). Structure and dynamics of the pan-genome of streptococcus pneumoniae and closely related species. *Genome Biol*, 11(10):R107.

Doolittle, W. F. (1973). Postmaturational cleavage of 23s ribosomal ribonucleic acid and its metabolic control in the blue-green alga anacystis nidulans. *J Bacteriol*, 113(3):1256–63.

Ducat, D. C., Way, J. C., and Silver, P. A. (2011). Engineering cyanobacteria to generate high-value products. *Trends Biotechnol*, 29(2):95–103.

Dutta, K., Daverey, A., and Lin, J.-G. (2014). Evolution retrospective for alternative fuels: First to fourth generation. *Renew Energy*, 69:114–122.

Dvornyk, V., Vinogradova, O., and Nevo, E. (2003). Origin and evolution of circadian clock genes in prokaryotes. *Proc Natl Acad Sci U S A*, 100(5):2495–500.

Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Minx, J. C., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., et al. (2015). Climate change 2014: mitigation of climate change. Technical report, Intergovernmental Panel on Climate Change.

Edgar, R. S., Green, E. W., Zhao, Y., van Ooijen, G., Olmedo, M., Qin, X., Xu, Y., Pan, M., Valekunja, U. K., Feeney, K. A., Maywood, E. S., Hastings, M. H., Baliga, N. S., Merrow, M., Millar, A. J., Johnson, C. H., Kyriacou, C. P., O'Neill, J. S., and Reddy, A. B. (2012). Peroxiredoxins are conserved markers of circadian rhythms. *Nat*, 485(7399):459–64.

Edwards, G. and Huber, S. (2014). The c4 pathway. In Hatch, N. and Boardman, N., editors, *The biochemistry of plants: a comprehensive treatise*, volume 8, chapter 6, pages 237–81. Elsevier.

Ellis, R. (1979). The most abundant protein in the world. *Trends Biochem Sci*, 4(11):241–4.

Enault, F., Suhre, K., Abergel, C., Poirot, O., and Claverie, J.-M. (2003). Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinforma*, 19 Suppl 1:i105–7.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A., and Martiny, A. C. (2013). Present and future global distributions of the marine cyanobacteria prochlorococcus and synechococcus. *Proc Natl Acad Sci U S A*, 110(24):9824–9.

Flores, F. G. and Herrero, A. (2008). *The Cyanobacteria: Molecular Biology, Genomics and Evolution*. Horizon Scientific Press.

Fouts, D. E., Brinkac, L., Beck, E., Inman, J., and Sutton, G. (2012). Panoct: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res*, 40(22):e172.

García-Fernández, J. M., de Marsac, N. T., and Diez, J. (2004). Streamlined regulation and gene loss as adaptive mechanisms in prochlorococcus for optimized nitrogen utilization in oligotrophic environments. *Microbiol Mol Biol Rev*, 68(4):630–8.

Garcia-Pichel, F., Belnap, J., Neuer, S., and Schanz, F. (2003). Estimates of global cyanobacterial biomass and its distribution. *Algol Stud*, 109(1):213–27.

Garcia-Pichel, F., López-Cortés, A., and Nübel, U. (2001). Phylogenetic and morphological diversity of cyanobacteria in soil desert crusts from the colorado plateau. *Appl Environ Microbiol*, 67(4):1902–10.

Georg, J. and Hess, W. R. (2011). cis-antisense rna, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*, 75(2):286–300.

Georg, J., Voss, B., Scholz, I., Mitschke, J., Wilde, A., and Hess, W. R. (2009). Evidence for a major role of antisense rnas in cyanobacterial gene regulation. *Mol Syst Biol*, 5:305.

Goericke, R. and Welschmeyer, N. A. (1993). The marine prochlorophyte prochlorococcus contributes significantly to phytoplankton biomass and primary production in the sargasso sea. *Deep Sea Res Part I: Oceanogr Res Pap*, 40(11):2283–94.

Gophna, U., Bapteste, E., Doolittle, W. F., Biran, D., and Ron, E. Z. (2005). Evolutionary plasticity of methionine biosynthesis. *Gene*, 355:48–57.

Grula, J. W. (2010). Rethinking the paleoproterozoic great oxidation event: A biological perspective. In *Astrobiology Science Conference 2010: Evolution and Life: Surviving Catastrophes and Extremes on Earth and Beyond*, volume 1538 of *LPI Contributions*, page 5071.

Gründel, M., Knoop, H., and Steuer, R. (2017). Activity and functional properties of the isocitrate lyase in the cyanobacterium cyanothece sp. pcc 7424. *Microbiol*, 163(5):731–44.

Guerreiro, A. C. L., Benevento, M., Lehmann, R., van Breukelen, B., Post, H., Giansanti, P., Maarten Altelaar, A. F., Axmann, I. M., and Heck, A. J. R. (2014). Daily rhythms in the cyanobacterium synechococcus elongatus probed by high-resolution mass spectrometry-based proteomics reveals a small defined set of cyclic proteins. *Mol Cell Proteomics*, 13(8):2042–55.

Gupta, R. S. and Mathews, D. W. (2010). Signature proteins for the major clades of cyanobacteria. *BMC Evol Biol*, 10:24.

Gupta, R. S., Pereira, M., Chandrasekera, C., and Johari, V. (2003). Molecular signatures in protein sequences that are characteristic of cyanobacteria and plastid homologues. *Int J Syst Evol Microbiol*, 53(Pt 6):1833–42.

Habib, M. A. B., Parvin, M., Huntington, T. C., and Hasan, M. R. (2008). *A review on culture, production and use of spirulina as food for humans and feeds for domestic animals and fish.* Food and agriculture organization of the united nations.

Hagemann, M. and Erdmann, N. (1994). Activation and pathway of glucosylglycerol synthesis in the cyanobacterium-synechocystis sp pcc-6803. *Microbiol*, 140:1427–31.

Hansen, M. C., Nielsen, A. K., Molin, S., Hammer, K., and Kilstrup, M. (2001). Changes in rrna levels during stress invalidates results from mrna blotting: fluorescence in situ rrna hybridization permits renormalization for estimation of cellular mrna levels. *J Bacteriol*, 183(16):4747–51.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., et al. (2004). The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61.

Hayashi, F., Suzuki, H., Iwase, R., Uzumaki, T., Miyake, A., Shen, J.-R., Imada, K., Furukawa, Y., Yonekura, K., Namba, K., and Ishiura, M. (2003). Atp-induced hexameric ring structure of the cyanobacterial circadian clock protein kaic. *Genes Cells*, 8(3):287–96.

Hekstra, D. and Tommassen, J. (1993). Functional exchangeability of the abc proteins of the periplasmic binding protein-dependent transport systems ugp and mal of escherichia coli. *J Bacteriol*, 175(20):6546–52.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol*, 28(9):977–82.

Hess, W. R. (2004). Genome analysis of marine photosynthetic microbes and their global role. *Curr Opin Biotechnol*, 15(3):191–8.

Higo, A., Katoh, H., Ohmori, K., Ikeuchi, M., and Ohmori, M. (2006). The role of a gene cluster for trehalose metabolism in dehydration tolerance of the filamentous cyanobacterium anabaena sp. pcc 7120. *Microbiol*, 152(Pt 4):979–87.

Hoffmann, L. (2003). *Caves and Other Low-Light Environments: Aerophitic Photoautotrophic Microorganisms*. John Wiley & Sons, Inc.

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C., and Ehrlich, G. D. (2007). Characterization and modeling of the haemophilus influenzae core and supragenomes based on the complete genomic sequences of rd and 12 clinical nontypeable strains. *Genome Biol*, 8(6):R103.

Holland, H. D. (2006). The oxygenation of the atmosphere and oceans. *Philos Transactions Royal Soc Lond B: Biol Sci*, 361(1470):903–15.

Holtzendorff, J., Partensky, F., Jacquet, S., Bruyant, F., Marie, D., Garczarek, L., Mary, I., Vaulot, D., and Hess, W. R. (2001). Diel expression of cell cycle-related genes in synchronized cultures of prochlorococcus sp. strain pcc 9511. *J Bacteriol*, 183(3):915–20.

Holtzendorff, J., Partensky, F., Mella, D., Lennon, J.-F., Hess, W. R., and Garczarek, L. (2008). Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium prochlorococcus marinus pcc 9511. *J Biol Rhythm*, 23(3):187–99.

Howitt, C. A. and Vermaas, W. F. (1998). Quinol and cytochrome oxidases in the cyanobacterium synechocystis sp. pcc 6803. *Biochem*, 37(51):17944–51.

Huynen, M. A., Dandekar, T., and Bork, P. (1999). Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol*, 7(7):281–91.

Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C. R., Tanabe, A., Golden, S. S., Johnson, C. H., and Kondo, T. (1998). Expression of a gene cluster kaiabc as a circadian feedback process in cyanobacteria. *Sci*, 281(5382):1519–23.

Ito, H., Mutsuda, M., Murayama, Y., Tomita, J., Hosokawa, N., Terauchi, K., Sugita, C., Sugita, M., Kondo, T., and Iwasaki, H. (2009). Cyanobacterial daily life with kai-based circadian and diurnal genome-wide transcriptional control in synechococcus elongatus. *Proc Natl Acad Sci*, 106(33):14168–73.

Ito, H. and Tanaka, A. (2011). Evolution of a divinyl chlorophyll-based photosystem in prochlorococcus. *Proc Natl Acad Sci United States Am*, 108(44):18014–19.

Iwasaki, H., Nishiwaki, T., Kitayama, Y., Nakajima, M., and Kondo, T. (2002). Kaia-stimulated kaic phosphorylation in circadian timing loops in cyanobacteria. *Proc Natl Acad Sci*, 99(24):15788–93.

Jawahar, R. (2015). Renewables 2015 global status report. Technical report, REN21: Renewable Energy Policy Network for the 21st century.

Jayroe, D. S. (2015). Stressed induced changes in karenia brevis ribosomal rna. Master's thesis, University of Southern Mississippi, USA.

Jeffery, C. J. (1999). Moonlighting proteins. *Trends Biochem Sci*, 24(1):8–11.

Jensen, P. E. and Leister, D. (2014). Chloroplast evolution, structure and functions. *F1000Prime Rep*, 6:40.

Johnson, C. H., Zhao, C., Xu, Y., and Mori, T. (2017). Timing the day: what makes bacterial clocks tick? *Nat Rev Microbiol*, 15(4):232–42.

Jones, J. G., Young, D. C., and DasSarma, S. (1991). Structure and organization of the gas vesicle gene cluster on the halobacterium halobium plasmid pnrc100. *Gene*, 102(1):117–22.

Jorquera, O., Kiperstok, A., Sales, E. A., Embiruçu, M., and Ghirardi, M. L. (2010). Comparative energy life-cycle analyses of microalgal biomass production in open ponds and photobioreactors. *Bioresour Technol*, 101(4):1406–13.

Jothi, R., Przytycka, T. M., and Aravind, L. (2007). Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinforma*, 8:173.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume 3, chapter 24, pages 21–132. Academic Press.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–4.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., et al. (1996). Sequence analysis of the genome of the unicellular cyanobacterium synechocystis sp. strain pcc6803. ii. sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res*, 3(3):109–36.

Kannenberg, E. L. and Poralla, K. (1999). Hopanoid biosynthesis and function in bacteria. *Naturwissenschaften*, 86(4):168–76.

Kasting, J. F. (2005). Methane and climate during the precambrian era. *Precambrian Res*, 137(3):119–29.

Kaufman, A. J. (2014). Early earth: Cyanobacteria at work. *Nat Geosci*, 7(4):253–4.

Keeley, J. E. and Rundel, P. W. (2003). Evolution of cam and c4 carbon-concentrating mechanisms. *Int J Plant Sci*, 164(S3):55–77.

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., and Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of prochlorococcus. *PLoS Genet*, 3(12):e231.

Khara, B., Menon, N., Levy, C., Mansell, D., Das, D., Marsh, E. N., Leys, D., and Scrutton, N. S. (2013). Production of propane and other short-chain alkanes by structure-based engineering of ligand specificity in aldehyde-deformylating oxygenase. *Chembiochem*, 14(10):1204–8.

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., and Church, G. M. (2006). Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinforma*, 7:177.

Kim, P. J. and Price, N. D. (2011). Genetic co-occurrence network across sequenced microbes. *PLoS Comput Biol*, 7(12):e1002340.

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J., and Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr Opin Biotechnol*, 23(4):617–23.

Kirchman, D. L. (2012). *Processes in Microbial Ecology*. Oxford University Press.

Kislyuk, A. O., Haegeman, B., Bergman, N. H., and Weitz, J. S. (2011). Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genom*, 12:32.

Kitayama, Y., Iwasaki, H., Nishiwaki, T., and Kondo, T. (2003). Kaib functions as an attenuator of kaic phosphorylation in the cyanobacterial circadian clock system. *EMBO J*, 22(9):2127–34.

Kitayama, Y., Nishiwaki, T., Terauchi, K., and Kondo, T. (2008). Dual kaic-based oscillations constitute the circadian system of cyanobacteria. *Genes Dev*, 22(11):1513–21.

Kjeldgaard, N. and Kurland, C. (1963). The distribution of soluble and ribosomal rna as a function of growth rate. *J Mol Biol*, 6(4):341–8.

Klähn, S., Baumgartner, D., Pfreundt, U., Voigt, K., Schön, V., Steglich, C., and Hess, W. R. (2014). Alkane biosynthesis genes in cyanobacteria and their transcriptional organization. *Front Bioeng Biotechnol*, 2:24.

Kluge, A. G. and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Syst Zool*, 18(1):1–32.

Knoop, H., Gründel, M., Zilliges, Y., Lehmann, R., Hoffmann, S., Lockau, W., and Steuer, R. (2013). Flux balance analysis of cyanobacterial metabolism: the metabolic network of synechocystis sp. pcc 6803. *PLoS Comput Biol*, 9(6):e1003081.

Knoop, H., Zilliges, Y., Lockau, W., and Steuer, R. (2010). The metabolic network of synechocystis sp. pcc 6803: systemic properties of autotrophic growth. *Plant Physiol*, 154(1):410–22.

Ko, C. H. and Takahashi, J. S. (2006). Molecular components of the mammalian circadian clock. *Hum Mol Genet*, 15 Spec No 2:R271–7.

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–42.

Kopp, R. E., Kirschvink, J. L., Hilburn, I. A., and Nash, C. Z. (2005). The paleoproterozoic snowball earth: A climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc Natl Acad Sci United States Am*, 102(32):11131–6.

Krehenbrink, M., Oppermann-Sanio, F.-B., and Steinbüchel, A. (2002). Evaluation of non-cyanobacterial genome sequences for occurrence of genes encoding proteins homologous to cyanophycin synthetase and cloning of an active cyanophycin synthetase from acinetobacter sp. strain dsm 587. *Arch Microbiol*, 177(5):371–80.

Kromkamp, J. (1987). Formation and functional significance of storage products in cyanobacteria. *New Zealand J Mar Freshw Res*, 21(3):457–465.

Kroneck, P. M. and Torres, M. E. S. (2015). *Sustaining Life on Planet Earth: Metalloenzymes Mastering Dioxygen and Other Chewy Gases*, volume 15. Springer.

Kucho, K.-i., Okamoto, K., Tsuchiya, Y., Nomura, S., Nango, M., Kanehisa, M., and Ishiura, M. (2005). Global analysis of circadian expression in the cyanobacterium synechocystis sp. strain pcc 6803. *J Bacteriol*, 187(6):2190–9.

Kumar, A., Suthers, P. F., and Maranas, C. D. (2012). Metrxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinforma*, 13:6.

Kumar, M., Kulshreshtha, J., and Singh, G. P. (2011). Growth and biopigment accumulation of cyanobacterium spirulina platensis at different light intensities and temperature. *Braz J Microbiol*, 42(3):1128–35.

Kushige, H., Kugenuma, H., Matsuoka, M., Ehira, S., Ohmori, M., and Iwasaki, H. (2013). Genome-wide and heterocyst-specific circadian gene expression in the filamentous cyanobacterium anabaena sp. strain pcc 7120. *J Bacteriol*, 195(6):1276–84.

Labiosa, R. G., Arrigo, K. R., Tu, C. J., Bhaya, D., Bay, S., Grossman, A. R., and Shrager, J. (2006). Examination of diel changes in global transcript accumulation in synechocystis (cyanobacteria). *J Phycol*, 42(3):622–36.

Lai, M. C. and Lan, E. I. (2015). Advances in metabolic engineering of cyanobacteria for photosynthetic biochemical production. *Metab*, 5(4):636.

Lapierre, P. and Gogarten, J. P. (2009). Estimating the size of the bacterial pangenome. *Trends Genet*, 25(3):107–10.

Latysheva, N., Junker, V. L., Palmer, W. J., Codd, G. A., and Barker, D. (2012). The evolution of nitrogen fixation in cyanobacteria. *Bioinforma*, 28(5):603–6.

Lehmann, R., Machné, R., Georg, J., Benary, M., Axmann, I. M., and Steuer, R. (2013). How cyanobacteria pose new problems to old methods: challenges in microarray time series analysis. *BMC Bioinforma*, 14(1):1–16.

Lepp and Schmidt (1998). Nucleic acid content of synechococcus spp. during growth in continuous light and light/dark cycles. *Arch Microbiol*, 170(3):201–7.

Lesser, M. P., Mazel, C. H., Gorbunov, M. Y., and Falkowski, P. G. (2004). Discovery of symbiotic nitrogen-fixing cyanobacteria in corals. *Sci*, 305(5686):997–1000.

Li, Y., Horsman, M., Wu, N., Lan, C. Q., and Dubois-Calero, N. (2008). Biofuels from microalgae. *Biotechnol Prog*, 24(4):815–820.

Liu, X. Q. and Yang, J. (2004). Bacterial thymidylate synthase with intein, group ii intron, and distinctive thyx motifs. *J Bacteriol*, 186(18):6316–9.

Liu, Y., Tsinoremas, N. F., Johnson, C. H., Lebedeva, N. V., Golden, S. S., Ishiura, M., and Kondo, T. (1995). Circadian orchestration of gene expression in cyanobacteria. *Genes & Dev*, 9(12):1469–78.

Lo, K., Hahne, F., Brinkman, R. R., and Gottardo, R. (2009). flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinforma*, 10:145.

Loza-Correa, M., Sahr, T., Rolando, M., Daniels, C., Petit, P., Skarina, T., Gomez Valero, L., Dervins-Ravault, D., Honoré, N., Savchenko, A., and Buchrieser, C. (2014). The legionella pneumophila kai operon is implicated in stress response and confers fitness in competitive environments. *Environ Microbiol*, 16(2):359–81.

Lum, K. K., Kim, J., and Lei, X. G. (2013). Dual potential of microalgae as a sustainable biofuel feedstock and animal feed. *J Animal Sci Biotechnol*, 4(1):1–7.

Ma, P., Mori, T., Zhao, C., Thiel, T., and Johnson, C. H. (2016). Evolution of kaic-dependent timekeepers: A proto-circadian timing mechanism confers adaptive fitness in the purple bacterium rhodopseudomonas palustris. *PLoS Genet*, 12(3):1–19.

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T., and Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol*, 35(1):81–9.

Malatinszky, D., Steuer, R., and Jones, P. R. (2017). A comprehensively curated genome-scale two-cell model for the heterocystous cyanobacterium anabaena sp. pcc 7120. *Plant Physiol*, 173(1):509–23.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nat*, 402(6757):83–6.

Marin, K., Zuther, E., Kerstan, T., Kunert, A., and Hagemann, M. (1998). The ggps gene from synechocystis sp. strain pcc 6803 encoding glucosyl-glycerol-phosphate synthase is involved in osmolyte synthesis. *J Bacteriol*, 180(18):4843–9.

Markou, G. and Georgakakis, D. (2011). Cultivation of filamentous cyanobacteria (blue-green algae) in agro-industrial wastes and wastewaters: a review. *Appl Energy*, 88(10):3389–401.

Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., Anderson, I., Billis, K., Varghese, N., Mavromatis, K., Pati, A., Ivanova, N. N., and Kyrpides, N. C. (2014). Img 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res*, 42(Database issue):D560–7.

Martins, J., Peixe, L., and Vasconcelos, V. M. (2011). Unraveling cyanobacteria ecology in wastewater treatment plants (wwtp). *Microb Ecol*, 62(2):241–56.

Mazouni, K., Domain, F., Cassier-Chauvat, C., and Chauvat, F. (2004). Molecular analysis of the key cytokinetic components of cyanobacteria: Ftsz, zipn and mincde. *Mol Microbiol*, 52(4):1145–58.

McFadden, G. I. (1999). Endosymbiosis and evolution of the plant cell. *Curr Opin Plant Biol*, 2(6):513–9.

Meng, Q., Zhang, Y., and Liu, X.-Q. (2007). Rare group i intron with insertion sequence element in a bacterial ribonucleotide reductase gene. *J Bacteriol*, 189(5):2150–4.

Mitschke, J., Georg, J., Scholz, I., Sharma, C. M., Dienst, D., Bantscheff, J., Voss, B., Steglich, C., Wilde, A., Vogel, J., and Hess, W. R. (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium synechocystis sp. pcc6803. *Proc Natl Acad Sci U S A*, 108(5):2124–9.

Möllers, K. B., Cannella, D., Jørgensen, H., and Frigaard, N.-U. (2014). Cyanobacterial biomass as carbohydrate and nutrient feedstock for bioethanol production by yeast fermentation. *Biotechnol For Biofuels*, 7(1):1.

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M., and Palsson, B. O. (2013). Genome-scale metabolic reconstructions of multiple escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A*, 110(50):20338–43.

Montagud, A., Navarro, E., Fernández de Córdoba, P., Urchueguía, J. F., and Patil, K. R. (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst Biol*, 4:156.

Moore, L. R., Rocap, G., and Chisholm, S. W. (1998). Physiology and molecular phylogeny of coexisting prochlorococcus ecotypes. *Nat*, 393(6684):464–7.

Moreno, J., Vargas, M. A., Rodríguez, H., Rivas, J., and Guerrero, M. G. (2003). Outdoor cultivation of a nitrogen-fixing marine cyanobacterium, anabaena sp. atcc 33047. *Biomol Eng*, 20(4-6):191–7.

Mori, T., Saveliev, S. V., Xu, Y., Stafford, W. F., Cox, M. M., Inman, R. B., and Johnson, C. H. (2002). Circadian clock protein kaic forms atp-dependent hexameric rings and binds dna. *Proc Natl Acad Sci U S A*, 99(26):17203–8.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35(Web Server issue):W182–5.

Mueller, T. J., Berla, B. M., Pakrasi, H. B., and Maranas, C. D. (2013). Rapid construction of metabolic models for a family of cyanobacteria using a multiple source annotation workflow. *BMC Syst Biol*, 7:142.

Mulec, J., Kosi, G., and Vrhovšek, D. (2008). Characterization of cave aerophytic algal communities and effects of irradiance levels on production of pigments. *J Cave Karst Stud*, 70(1):3–12.

Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R., and Galperin, M. Y. (2006). The cyanobacterial genome core and the origin of photosynthesis. *Proc Natl Acad Sci U S A*, 103(35):13126–31.

Müller, M., Mentel, M., van Hellemond, J. J., Henze, K., Woehle, C., Gould, S. B., Yu, R.-Y., van der Giezen, M., Tielens, A. G. M., and Martin, W. F. (2012). Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev*, 76(2):444–95.

Mullineaux, C. W. and Stanewsky, R. (2009). The rolex and the hourglass: a simplified circadian clock in prochlorococcus? *J Bacteriol*, 191(17):5333–5.

Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., and Kondo, T. (2005). Reconstitution of circadian oscillation of cyanobacterial kaic phosphorylation in vitro. *Sci*, 308(5720):414–5.

Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Takeuchi, C., Yamada, M., and Tabata, S. (2003). Complete genome structure of gloeobacter violaceus pcc 7421, a cyanobacterium that lacks thylakoids. *DNA Res*, 10(4):137–45.

Nakamura, Y., Takahashi, J.-I., Sakurai, A., Inaba, Y., Suzuki, E., Nihei, S., Fujiwara, S., Tsuzuki, M., Miyashita, H., Ikemoto, H., Kawachi, M., Sekiguchi, H., and Kurano, N. (2005). Some cyanobacteria synthesize semi-amylopectin type alpha-polyglucans instead of glycogen. *Plant Cell Physiol*, 46(3):539–45.

Nishiwaki, T., Iwasaki, H., Ishiura, M., and Kondo, T. (2000). Nucleotide binding and autophosphorylation of the clock protein kaic as a circadian timing process of cyanobacteria. *Proc Natl Acad Sci U S A*, 97(1):495–9.

Nishiwaki, T. and Kondo, T. (2012). Circadian autodephosphorylation of cyanobacterial clock protein kaic occurs via formation of atp as intermediate. *J Biol Chem*, 287(22):18030–5.

Notebaart, R. A., van Enckevort, F. H., Francke, C., Siezen, R. J., and Teusink, B. (2006). Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinforma*, 7:296.

Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:320.

Ochoa de Alda, J. A. G., Esteban, R., Diago, M. L., and Houmard, J. (2014). The plastid ancestor originated among one of the major cyanobacterial lineages. *Nat Commun*, 5:4937.

Oldham, M. L., Davidson, A. L., and Chen, J. (2008). Structural insights into abc transporter mechanism. *Curr Opin Struct Biol*, 18(6):726–33.

Oliver, N. J., Rabinovitch-Deere, C. A., Carroll, A. L., Nozzi, N. E., Case, A. E., and Atsumi, S. (2016). Cyanobacterial metabolic engineering for biofuel and chemical production. *Curr Opin Chem Biol*, 35:43–50.

Oppermann-Sanio, F. B. and Steinbüchel, A. (2002). Occurrence, functions and biosynthesis of polyamides in microorganisms and biotechnological production. *Naturwissenschaften*, 89(1):11–22.

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat Biotechnol*, 28(3):245–8.

Ouyang, Y., Andersson, C. R., Kondo, T., Golden, S. S., and Johnson, C. H. (1998). Resonating circadian clocks enhance fitness in cyanobacteria. *Proc Natl Acad Sci U S A*, 95(15):8660–4.

Ovando, C. A., de Carvalho, J. C., de Melo Pereira, G. V., Jacques, P., Soccol, V. T., and Soccol, C. R. (2016). Functional properties and health benefits of bioactive peptides derived from spirulina: A review. *Food Rev Int*, pages 1–18.

Panaro, N. J., Yuen, P. K., Sakazume, T., Fortina, P., Kricka, L. J., and Wilding, P. (2000). Evaluation of dna fragment sizing and quantification by the agilent 2100 bioanalyzer. *Clin Chem*, 46(11):1851–3.

Parmar, A., Singh, N. K., Pandey, A., Gnansounou, E., and Madamwar, D. (2011). Cyanobacteria and microalgae: a positive prospect for biofuels. *Bioresour Technol*, 102(22):10163–72.

Pattanayek, R., Williams, D. R., Pattanayek, S., Mori, T., Johnson, C. H., Stewart, P. L., and Egli, M. (2008). Structural model of the circadian clock kaib-kaic complex and mechanism for modulation of kaic phosphorylation. *EMBO J*, 27(12):1767–78.

Pearce, J., Leach, C. K., and Carr, N. G. (1969). The incomplete tricarboxylic acid cycle in the blue-green alga anabaena variabilis. *J Gen Microbiol*, 55(3):371–8.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–8.

Penn, K., Wang, J., Fernando, S. C., and Thompson, J. R. (2014). Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. *ISME J*, 8(9):1866–78.

Peschek, G. A., Obinger, C., and Renger, G. (2011). *Bioenergetic Processes of Cyanobacteria*. Springer.

Philip, S., Keshavarz, T., and Roy, I. (2007). Polyhydroxyalkanoates: biodegradable polymers with a range of applications. *J Chem Technol & Biotechnol*, 82(3):233–247.

Pienkos, P. T. and Darzins, A. (2009). The promise and challenges of microalgal-derived biofuels. *Biofuels, Bioprod Biorefining*, 3(4):431–40.

Pisciotta, J. M., Zou, Y., and Baskakov, I. V. (2010). Light-dependent electrogenic activity of cyanobacteria. *PLoS One*, 5(5):e10821.

Poulsen, L. K., Ballard, G., and Stahl, D. A. (1993). Use of rrna fluorescence in situ hybridization for measuring the activity of single cells in young and established biofilms. *Appl Environ Microbiol*, 59(5):1354–60.

Quintana, N., Van der Kooy, F., Van de Rhee, M. D., Voshol, G. P., and Verpoorte, R. (2011). Renewable energy from cyanobacteria: energy production optimization by metabolic pathway engineering. *Appl Microbiol Biotechnol*, 91(3):471–90.

Rae, B. D., Long, B. M., Whitehead, L. F., Förster, B., Badger, M. R., and Price, G. D. (2013). Cyanobacterial carboxysomes: microcompartments that facilitate co2 fixation. *J Mol Microbiol Biotechnol*, 23(4-5):300–7.

Rai, A. N., Bergman, B., and Rasmussen, U. (2002). *Cyanobacteria in Symbiosis*. Springer.

Raoof, B., Kaushik, B., and Prasanna, R. (2006). Formulation of a low-cost medium for mass production of spirulina. *Biomass Bioenergy*, 30(6):537–42.

Raposo, M. F. d. J., de Morais, R. M. S. C., and Bernardo de Morais, A. M. M. (2013a). Bioactivity and applications of sulphated polysaccharides from marine microalgae. *Mar Drugs*, 11(1):233–52.

Raposo, M. F. d. J., de Morais, R. M. S. C., and de Morais, A. M. M. B. (2013b). Health applications of bioactive compounds from marine microalgae. *Life Sci*, 93(15):479–86.

Raven, J. (1991). Physiology of inorganic c acquisition and implications for resource use efficiency by marine phytoplankton: relation to increased co2 and temperature. *Plant, Cell & Environ*, 14(8):779–94.

Raven, J. A. (2009). Contributions of anoxygenic and oxygenic phototrophy and chemolithotrophy to carbon and oxygen fluxes in aquatic environments. *Aquatic Microb Ecol*, 56(2-3):177–92.

Raven, J. A. (2013). Rubisco: still the most abundant protein of earth? *New Phytol*, 198(1):1–3.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., and Blankenship, R. E. (2002). Whole-genome analysis of photosynthetic prokaryotes. *Sci*, 298(5598):1616–20.

Riccardi, G., de Rossi, E., and Milano, A. (1989). Amino acid biosynthesis and its regulation in cyanobacteria. *Plant Sci*, 64(2):135–51.

Rippka, R., Deruelles, J., Waterbury, J. B., Herdman, M., and Stanier, R. Y. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol*, 111(Mar):1–61.

Romano, A. H. and Conway, T. (1996). Evolution of carbohydrate metabolic pathways. *Res Microbiol*, 147(6-7):448–55.

Rothschild, L. and Lister, A. (2003). *Evolution on Planet Earth: Impact of the Physical Environment*. Academic Press.

Ruano-Rubio, V., Poch, O., and Thompson, J. D. (2009). Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC Bioinforma*, 10:383.

Rubio, L. M., Flores, E., and Herrero, A. (1999). Molybdopterin guanine dinucleotide cofactor in synechococcus sp. nitrate reductase: identification of moba and isolation of a putative moeb gene. *FEBS Lett*, 462(3):358–62.

Rust, M. J., Golden, S. S., and O'Shea, E. K. (2011). Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator. *Sci*, 331(6014):220–3.

Saha, R., Verseput, A. T., Berla, B. M., Mueller, T. J., Pakrasi, H. B., and Maranas, C. D. (2012). Reconstruction and comparison of the metabolic potential of cyanobacteria cyanothece sp. atcc 51142 and synechocystis sp. pcc 6803. *PLoS One*, 7(10):e48285.

Savakis, P. and Hellingwerf, K. J. (2015). Engineering cyanobacteria for direct biofuel production from co 2. *Curr Opin Biotechnol*, 33:8–14.

Savakis, P., Tan, X., Qiao, C., Song, K., Lu, X., Hellingwerf, K. J., and Branco Dos Santos, F. (2016). Slr1670 from synechocystis sp. pcc 6803 is required for the re-assimilation of the osmolyte glucosylglycerol. *Front Microbiol*, 7:1350.

Schirmer, A., Rude, M. A., Li, X., Popova, E., and del Cardayre, S. B. (2010). Microbial biosynthesis of alkanes. *Sci*, 329(5991):559–62.

Schmelling, N. M., Lehmann, R., Chaudhury, P., Beck, C., Albers, S.-V., Axmann, I. M., and Wiegard, A. (2017). Minimal tool set for a prokaryotic circadian clock. *BMC Evol Biol*, 17(1):169.

Schnarrenberger, C. and Martin, W. (2002). Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants. *Eur J Biochem*, 269(3):868–83.

Seckbach, J. (2007). *Algae and Cyanobacteria in Extreme Environments*, volume 11. Springer Science & Business Media.

Sessions, A. L., Doughty, D. M., Welander, P. V., Summons, R. E., and Newman, D. K. (2009). The continuing puzzle of the great oxidation event. *Curr Biol*, 19(14):R567–74.

Sherman, L. A., Meunier, P., and Colón-López, M. S. (1998). Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynth Res*, 58(1):25–42.

Shi, T. and Falkowski, P. G. (2008). Genome evolution in cyanobacteria: the stable core and the variable shell. *Proc Natl Acad Sci U S A*, 105(7):2510–5.

Shi, T., Ilikchyan, I., Rabouille, S., and Zehr, J. P. (2010). Genome-wide analysis of diel gene expression in the unicellular n(2)-fixing cyanobacterium crocosphaera watsonii wh 8501. *ISME J*, 4(5):621–32.

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K. W., Han, C. S., Rubin, E. M., Eisen, J. A., Woyke, T., Gugger, M., and Kerfeld, C. A. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A*, 110(3):1053–8.

Simm, S., Keller, M., Selymesi, M., and Schleiff, E. (2015). The composition of the global and feature specific cyanobacterial core-genomes. *Front Microbiol*, 6:219.

Simonsen, M., Maetschke, S. R., and Ragan, M. A. (2012). Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling. *Bioinforma*, 28(6):851–7.

Singh, R., Parihar, P., Singh, M., Bajguz, A., Kumar, J., Singh, S., Singh, V. P., and Prasad, S. M. (2017). Uncovering potential applications of cyanobacteria and algal metabolites in biology, agriculture and medicine: Current status and future prospects. *Front Microbiol*, 8:515.

Sinha, R. P. and Häder, D.-P. (1996). Photobiology and ecophysiology of rice field cyanobacteria. *Photochem Photobiol*, 64(6):887–96.

Škunca, N. and Dessimoz, C. (2015). Phylogenetic profiling: how much input data is enough? *PLoS One*, 10(2):e0114701.

Smith, A. J., London, J., and Stanier, R. Y. (1967). Biochemical basis of obligate autotrophy in blue-green algae and thiobacilli. *J Bacteriol*, 94(4):972–83.

Smyth, G. K. (2005). *limma: Linear Models for Microarray Data*, pages 397–420. Springer New York, NY.

Snitkin, E. S., Gustafson, A. M., Mellor, J., Wu, J., and DeLisi, C. (2006). Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinforma*, 7:420.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–38.

Sompong, U., Hawkins, P. R., Besley, C., and Peerapornpisal, Y. (2005). The distribution of cyanobacteria across physical and chemical gradients in hot springs in northern thailand. *FEMS Microbiol Ecol*, 52(3):365–76.

Sousa, F. L., Shavit-Grievink, L., Allen, J. F., and Martin, W. F. (2013). Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. *Genome Biol Evol*, 5(1):200–16.

Stal, L. J. and Moezelaar, R. (1997). Fermentation in cyanobacteria. *FEMS Microbiol Rev*, 21(2):179–211.

Steinbüchel, A., Füchtenbusch, B., Gorenflo, V., Hein, S., Jossek, R., Langenbach, S., and Rehm, B. H. (1998). Biosynthesis of polyesters in bacteria and recombinant organisms. *Polym Degrad Stab*, 59(1):177–82.

Steinhauser, D., Fernie, A. R., and Araújo, W. L. (2012). Unusual cyanobacterial tca cycles: not broken just different. *Trends Plant Sci*, 17(9):503–9.

Steuer, R., Knoop, H., and Machné, R. (2012). Modelling cyanobacteria: from metabolism to integrative models of phototrophic growth. *J Exp Bot*, 63(6):2259–74.

Stöckel, J., Welsh, E. A., Liberton, M., Kunnvakkam, R., Aurora, R., and Pakrasi, H. B. (2008). Global transcriptomic analysis of cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc Natl Acad Sci U S A*, 105(16):6156–61.

Straub, C., Quillardet, P., Vergalli, J., de Marsac, N. T., and Humbert, J.-F. (2011). A day in the life of microcystis aeruginosa strain pcc 7806 as revealed by a transcriptomic analysis. *PLoS ONE*, 6(1):1–12.

Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A., Shi, T., and Li, Y. (2005). Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinforma*, 21(16):3409–15.

Suutari, M., Majaneva, M., Fewer, D. P., Voirin, B., Aiello, A., Friedl, T., Chiarello, A. G., and Blomster, J. (2010). Molecular evidence for a diverse green algal community growing in the hair of sloths and a specific association with trichophilus welckeri (chlorophyta, ulvophyceae). *BMC Evol Biol*, 10:86.

Tabita, F. R. (2004). The biochemistry and molecular regulation of carbon dioxide metabolism in cyanobacteria. In Bryant, D. A., editor, *The molecular biology of cyanobacteria*, volume 1 of *Advances in Photosynthesis*, chapter 14, pages 437–67. Springer Science & Business Media.

Takaichi, S. and Mochimaru, M. (2007). Carotenoids and carotenogenesis in cyanobacteria: unique ketocarotenoids and carotenoid glycosides. *Cell Mol Life Sci*, 64(19-20):2607–19.

Tanioka, Y., Yabuta, Y., Yamaji, R., Shigeoka, S., Nakano, Y., Watanabe, F., and Inui, H. (2009). Occurrence of pseudovitamin b12 and its possible function as the cofactor of cobalamin-dependent methionine synthase in a cyanobacterium synechocystis sp. pcc6803. *J Nutr Sci Vitaminol (Tokyo)*, 55(6):518–21.

Tcherkez, G. G. B., Farquhar, G. D., and Andrews, T. J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc Natl Acad Sci U S A*, 103(19):7246–51.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., et al. (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*, 102(39):13950–5.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*, 11(5):472–7.

Thiele, I. and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121.

Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bolling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., et al. (2013). A community-driven global reconstruction of human metabolism. *Nat Biotechnol*, 31(5):419–25.

Thompson, A. W., Foster, R. A., Krupke, A., Carter, B. J., Musat, N., Vaulot, D., Kuypers, M. M. M., and Zehr, J. P. (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Sci*, 337(6101):1546–50.

Tian, Y., Zhang, J., Song, L., and Bao, H. (2001). A study on aerial cyanophyta (cyanobacteria) on the surface of carbonate rock in yunnan stone forest, yunnan province, china. *Acta Ecol Sinica*, 22(11):1793–802.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*, 5(2):123–35.

Ting, I. P. (1985). Crassulacean acid metabolism. *Annu Rev Plant Physiol*, 36(1):595–622.

Toepel, J., Welsh, E., Summerfield, T. C., Pakrasi, H. B., and Sherman, L. A. (2008). Differential transcriptional analysis of the cyanobacterium cyanothece sp. strain atcc 51142 during light-dark and continuous-light growth. *J Bacteriol*, 190(11):3904–13.

Toepel, J. R., McDermott, J. E., Summerfield, T. C., and Sherman, L. A. (2009). Transcriptional analysis of the unicellular, diazotrophic cyanobacterium cyanothece sp. atcc 51142 grown under short day/night cycles. *J Phycol*, 45(3):610–20.

Tomii, K. and Kanehisa, M. (1998). A comparative analysis of abc transporters in complete microbial genomes. *Genome Res*, 8(10):1048–59.

Triana, J., Montagud, A., Siurana, M., Fuente, D., Urchueguía, A., Gamermann, D., Torres, J., Tena, J., de Córdoba, P. F., and Urchueguía, J. F. (2014). Generation and evaluation of a genome-scale metabolic network model of synechococcus elongatus pcc7942. *Metab*, 4(3):680–98.

Turner, J. F. and Turner, D. H. (1980). *The Regulation of Glycolysis and the Pentose Phosphate Pathway*, volume 2 of *The Biochemistry of Plants: a comprehensive treatise*, chapter 7, pages 279–316. Academic Press, metabolisms and respiration edition.

Ungerer, J., Tao, L., Davis, M., Ghirardi, M., Maness, P.-C., and Yu, J. (2012). Sustained photosynthetic conversion of co 2 to ethylene in recombinant cyanobacterium synechocystis 6803. *Energy & Environ Sci*, 5(10):8998–9006.

van Alphen, P. and Hellingwerf, K. J. (2015). Sustained circadian rhythms in continuous light in synechocystis sp. pcc6803 growing in a well-controlled photobioreactor. *PLOS ONE*, 10(6):1–12.

Vert, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinforma*, 18 Suppl 1:S276–84.

Vieira, G., Sabarly, V., Bourguignon, P. Y., Durot, M., Le Fevre, F., Mornico, D., Vallenet, D., Bouvet, O., Denamur, E., Schachter, V., and Medigue, C. (2011). Core and panmetabolism in escherichia coli. *J Bacteriol*, 193(6):1461–72.

Vijayan, V., Zuzow, R., and O'Shea, E. K. (2009). Oscillations in supercoiling drive circadian gene expression in cyanobacteria. *Proc Natl Acad Sci*, 106(52):22564–8.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J Mach Learn Res*, 11:2837–54.

Vitkin, E. and Shlomi, T. (2012). Mirage: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biol*, 13(11):R111.

Volpe, J. J. and Vagelos, P. R. (1976). Mechanisms and regulation of biosynthesis of saturated fatty acids. *Physiol Rev*, 56(2):339–417.

Vu, T. T., Stolyar, S. M., Pinchuk, G. E., Hill, E. A., Kucek, L. A., Brown, R. N., Lipton, M. S., Osterman, A., Fredrickson, J. K., Konopka, A. E., Beliaev, A. S., and Reed, J. L. (2012). Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium *Cyanothece* sp. atcc 51142. *PLoS Comput Biol*, 8(4):1–15.

Wang, B., Pugh, S., Nielsen, D. R., Zhang, W., and Meldrum, D. R. (2013). Engineering cyanobacteria for photosynthetic production of 3-hydroxybutyrate directly from co 2. *Metab Eng*, 16:68–77.

Wang, B., Wang, J., Zhang, W., and Meldrum, D. R. (2012). Application of synthetic biology in cyanobacteria and algae. *Front Microbiol*, 3:344.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

Warren-Rhodes, K. A., Rhodes, K. L., Pointing, S. B., Ewing, S. A., Lacap, D. C., Gómez-Silva, B., Amundson, R., Friedmann, E. I., and McKay, C. P. (2006). Hypolithic cyanobacteria, dry limit of photosynthesis, and microbial ecology in the hyperarid atacama desert. *Microb Ecol*, 52(3):389–98.

Welkie, D., Zhang, X., Markillie, M. L., Taylor, R., Orr, G., Jacobs, J., Bhide, K., Thimmapuram, J., Gritsenko, M., Mitchell, H., Smith, R. D., and Sherman, L. A. (2014). Transcriptomic and proteomic dynamics in the metabolism of a diazotrophic cyanobacterium, cyanothece sp. pcc 7822 during a diurnal light–dark cycle. *BMC Genom*, 15(1):1–16.

Welp, L. R., Keeling, R. F., Meijer, H. A. J., Bollenbacher, A. F., Piper, S. C., Yoshimura, K., Francey, R. J., Allison, C. E., and Wahlen, M. (2011). Interannual variability in the oxygen isotopes of atmospheric co2 driven by el niño. *Nat*, 477(7366):579–82.

Whitehead, L., Long, B. M., Price, G. D., and Badger, M. R. (2014). Comparing the in vivo function of alpha-carboxysomes and beta-carboxysomes in two model cyanobacteria. *Plant Physiol*, 165(1):398–411.

Whitton, B. A. (2012). *Ecology of cyanobacteria II: their diversity in space and time.* Springer Science & Business Media.

Wiegard, A., Dörrich, A. K., Deinzer, H.-T., Beck, C., Wilde, A., Holtzendorff, J., and Axmann, I. M. (2013). Biochemical analysis of three putative kaic clock proteins from synechocystis sp. pcc 6803 suggests their functional divergence. *Microbiol*, 159(5):948–58.

Winter, S., Jahn, K., Wehner, S., Kuchenbecker, L., Marz, M., Stoye, J., and Bocker, S. (2016). Finding approximate gene clusters with gecko 3. *Nucleic Acids Res*, 44(20):9600–10.

Woelfle, M. A. and Johnson, C. H. (2009). The adaptive value of the circadian clock system in cyanobacteria. In *Bacterial Circadian Programs*, chapter 12, pages 205–21. Springer.

Woelfle, M. A., Ouyang, Y., Phanvijhitsiri, K., and Johnson, C. H. (2004). The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. *Curr Biol*, 14(16):1481–6.

Woo, H. M. and Lee, H. J. (2017). Toward solar biodiesel production from co2 using engineered cyanobacteria. *FEMS Microbiol Lett*, 364(9).

Woodard, L. M., Bielkie, A. R., Eisses, J. F., and Ketchum, P. A. (1990). Occurrence of nitrate reductase and molybdopterin in xanthomonas maltophilia. *Appl Environ Microbiol*, 56(12):3766–71.

Woodrow, I. E. and Berry, J. A. (1988). Enzymatic regulation of photosynthetic co2, fixation in c3 plants. *Annu Rev Plant Physiol Plant Mol Biol*, 39(1):533–94.

Wu, J., Kasif, S., and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. *Bioinforma*, 19(12):1524–30.

Wulff, J. L. (2006). Ecological interactions of marine sponges. *Can J Zool*, 84(2):146–66.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., and Johnson, C. H. (2013a). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nat*, 495(7439):116–20.

Xu, Y.-F., Létisse, F., Absalan, F., Lu, W., Kuznetsova, E., Brown, G., Caudy, A. A., Yakunin, A. F., Broach, J. R., and Rabinowitz, J. D. (2013b). Nucleotide degradation and ribose salvage in yeast. *Mol Syst Biol*, 9:665.

Yoshikawa, K., Aikawa, S., Kojima, Y., Toya, Y., Furusawa, C., Kondo, A., and Shimizu, H. (2015a). Construction of a genome-scale metabolic model of arthrospira platensis nies-39 and metabolic design for cyanobacterial bioproduction. *PLoS One*, 10(12):e0144430.

Yoshikawa, K., Hirasawa, T., and Shimizu, H. (2015b). Effect of malic enzyme on ethanol production by synechocystis sp. pcc 6803. *J Biosci Bioeng*, 119(1):82–4.

Zakhia, F., Jungblut, A.-D., Taton, A., Vincent, W. F., and Wilmotte, A. (2008). Cyanobacteria in cold ecosystems. In *Psychrophiles: from biodiversity to biotechnology*, chapter 8, pages 121–35. Springer.

Zehr, J. P. (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol*, 19(4):162–73.

Zehr, J. P., Bench, S. R., Carter, B. J., Hewson, I., Niazi, F., Shi, T., Tripp, H. J., and Affourtit, J. P. (2008). Globally distributed uncultivated oceanic n2-fixing cyanobacteria lack oxygenic photosystem ii. *Sci*, 322(5904):1110–2.

Zhang, C. C., Jeanjean, R., and Joset, F. (1998). Obligate phototrophy in cyanobacteria: more than a lack of sugar transport. *FEMS Microbiol Lett*, 161(2):285–92.

Zhang, S. and Bryant, D. A. (2011). The tricarboxylic acid cycle in cyanobacteria. *Sci*, 334(6062):1551–3.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res*, 16(9):1099–108.

Zheng, W., Bergman, B., Chen, B., Zheng, S., Guan, X., and Rasmussen, U. (2009). Cellular responses in the cyanobacterial symbiont during its vertical transfer between plant generations in the azolla microphylla symbiosis. *New Phytol*, 181(1):53–61.

Zhou, Y., Wang, R., Li, L., Xia, X., and Sun, Z. (2006). Inferring functional linkages between proteins from evolutionary scenarios. *J Mol Biol*, 359(4):1150–9.

Zinser, E. R., Lindell, D., Johnson, Z. I., Futschik, M. E., Steglich, C., Coleman, M. L., Wright, M. A., Rector, T., Steen, R., McNulty, N., Thompson, L. R., and Chisholm, S. W. (2009). Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, prochlorococcus. *PLoS ONE*, 4(4):1–18.

List of Figures

List of Tables

List of Abbreviations

asRNA	(*cis*-)antisense RNA
ATP	adenosine triphosphate
BHR	bidirectional hit rate
BLAST	basic local alignment search tool
BLASTp	protein-protein BLAST
bp	base pair
CLOG	cluster of likely orthologous genes
DNA	deoxyribonucleic acid
DFT	discrete Fourier transform
EC number	Enzyme Commission number
FBA	flux balance analysis
G+C	percentage of guanine and cytosine
GO	gene ontology
LAHG	light activated heterotrophic growth
LL	continuous light conditions
LD	oscillating light-dark conditions
LDLDL	oscillating light-dark conditions in 48h experiment
LDLLL	48h experiment with transfer to continuous light
LDDDD	48h experiment with transfer to continuous dark
LOS	least oscillatory set
NAD	nicotinamide adenine dinucleotide
NADP	nicotinamide adenine dinucleotide phosphate
ncRNA	small non-coding RNA
nt	nucleotide
ORF	open reading frame
PEP	phosphoenolpyruvate
PHB	poly-β-hydroxybutyrate
R	programming language for statistical computing
RNA	ribonucleic acid
rRNA	ribosomal RNA
TCA cycle	tricarboxylic acid cycle
tRNA	transfer RNA
5'-UTR	regulatory untranslated region upstream of a gene initiation codon

List of publications

Peer-reviewed journal articles

Beck C, Knoop H, Axmann IM, and Steuer R (2012). The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics*, 13(1):1–17.

Wiegard A, Dörrich AK, Deinzer H-T, Beck C, Wilde A, Holtzendorff J, and Axmann IM (2013). Biochemical analysis of three putative KaiC clock proteins from Synechocystis sp. PCC 6803 suggests their functional divergence. *Microbiology*, 159(5):948–58.

Beck C, Hertel S, Rediger A, Lehmann R, Wiegard A, Kölsch A, Heilmann B, Georg J, Hess WR, and Axmann IM (2014). Daily expression pattern of protein-encoding genes and small noncoding RNAs in Synechocystis sp. strain PCC 6803. *Applied and Environmental Microbiology*, 80(17):5195–206.

Schmelling NM, Lehmann R, Chaudhury P, Beck C, Albers S-V, Axmann IM, and Wiegard A (2017). Minimal tool set for a prokaryotic circadian clock. *BMC Evolutionary Biology*, 17(1):169.

Beck C, Knoop H, and Steuer R (submitted). Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between ortholog genes. Preprint: doi.org/10.1101/137398.

Oral presentation

A daily temporal program for rhythmic expression in Synechocystis PCC 6803. *12th Workshop on Cyanobacteria* 2013, St. Louis, MO, USA.

Poster presentations (excerpt)

Evidence for a less robust timing mechanism in Prochlorococcus. *3rd FEBS Advanced Lecture Course of Systems Biology: From Molecular Biology to Biological Function* 2009, Alpbach, Austria.

The diversity of cyanobacterial metabolism: A multi genome comparison. *8th European Workshop on Molecular Biology of Cyanobacteria* 2011, Naantali, Finland.

Appendices

Appendix A.

General information for cyanobacterial strains considered in this thesis

In this table, we provide genomic and growth information for each strain including natural habitat, morphology (sections I-V, according to Rippka et al. 1979), number of chromosomes & plasmids, number of ORFs, genome size (in megabase pairs), G+C content (in percent), fraction of DNA in ORFs (in percent), number of CLOGs, number of core CLOGs, number of shared CLOGs, number of unique CLOGs, and number of CLOGs with assigned metabolic function. The number of CLOGs were taken from the study comparing all 78 genomes, presented in Chapter 4. We also extracted from literature the strains' ability to fixate atmospheric nitrogen. Literature data disagreeing with the findings in our study (strain has no orthologs in module 9, composed of CLOGs mostly associated to nitrogenase) is marked with an asterisk. The last column lists various information concerning habitat, metabolism, symbiosis, and particular features of the strains. Organisms of the genus *Prochlorococcus* are annotated with the water depth at which the according strain was found, and their adaptation to high light (HL) or low light (LL). If not noted otherwise, data regarding the structural section was extracted from Shih et al. 2013, while information regarding habitat, nitrogen fixation, and general properties was extracted from Markowitz et al. 2014. The 16 cyanobacterial strains considered in the study presented in Chapter 3 are highlighted in bold.

Name	Habitat[5]	Morphological sections[12]	Chromosomes (plasmids)	ORFs	Genome size (Mb)	G+C content (%)	DNA coding (%)	CLOGs	Core CLOGs	Shared CLOGs	Unique CLOGs	Metabolic CLOGs	Nitrogen fixation[5]	Adaptation[5]
***Acaryochloris marina* MBIC11017**	marine	I	1 (9)	8383	8.36	46.96	82.46	7662	620	4120	2922	3074	no	
Anabaena cylindrica PCC 7122	fresh water	IV	1 (6)	5838	7.06	38.79	79.98	5279	620	4234	425	1324	yes	motile
Anabaena sp. 90	fresh water	IV[1]	2 (3)	4511	5.31	38.1	79.78	4161	620	3104	437	1177	yes	motile
Anabaena variabilis ATCC 29413 *(Anabaena flos-aquae* UTEX 1444)	fresh water, soil	IV	2 (3)	5706	7.11	41.41	82.05	5070	620	4227	223	1271	yes	motile
Arthrospira platensis NIES-39	fresh water	III	1 (0)	6630	6.79	44.27	81.22	6239	620	3234	2385	2412	no	
Calothrix sp. 336/3	fresh water	IV[2]	4 (0)	4834	6.42	41.11	71.41	4385	620	3436	329	1238	no*	H_2 producing[2]
Calothrix sp. PCC 6303	fresh water	IV	1 (3)	5535	6.96	39.8	79	5001	620	3898	483	1318	yes	photoheterotroph
Calothrix sp. PCC 7507	fresh water	IV	1 (0)	5950	7.02	42.25	79.31	5326	620	4210	496	1416	yes	
Chamaesiphon minutus PCC 6605	fresh water	I	1 (2)	5945	6.76	45.67	79.68	5436	620	3616	1200	1711	no	
Chroococcidiopsis thermalis PCC 7203	terrestrial, soil	II	1 (2)	5752	6.69	44.47	82.66	5077	620	3892	565	1473	yes	non-motile, aerobe, heterotroph
Crinalium epipsammum PCC 9333	terrestrial[3]	III	1 (8)	4752	5.62	40.16	81.04	4383	620	3234	529	1323	no	sand dunes, drought tolerant
Cyanobacterium aponinum PCC 10605	fresh water	I	1 (1)	3431	4.18	34.93	80.09	3208	620	2360	228	1023	yes*	
Cyanobacterium stanieri PCC 7202	thermophilic, alkaline	I	1 (0)	2837	3.16	38.66	85.84	2661	620	1913	128	895	no	
Cyanobium gracile PCC 6307	fresh water	I	1 (0)	3280	3.34	68.71	89.28	3087	620	1984	483	1153	no	nonmotile, aerobe
***Cyanothece* sp. ATCC 51142**	marine, intertidal	I	2 (4)	5304	5.46	37.94	86.31	4842	620	3463	759	1479	yes	nonmotile, photoheterotroph
Cyanothece sp. PCC 7424	fresh water, terrestrial, rice field	I	1 (6)	5710	6.55	38.51	81.1	5159	620	4048	491	1358	yes	nonmotile, photoheterotroph, anaerobe
Cyanothece sp. PCC 7425	fresh water, terrestrial, rice field	I	1 (3)	5327	5.79	50.65	85.04	4845	620	3475	750	1483	yes	nonmotile, photoheterotroph, anaerobe
Cyanothece sp. PCC 7822	fresh water, terrestrial, rice field	I	1 (6)	6642	7.84	39.9	82.59	6007	620	4517	870	1570	yes	nonmotile, aerobe
***Cyanothece* sp. PCC 8801**	fresh water, terrestrial, rice field	I	1 (3)	4367	4.79	39.76	84.54	4026	620	3334	72	1046	yes	nonmotile, aerobe
Cyanothece sp. PCC 8802	fresh water, terrestrial, rice field	I	1 (4)	4444	4.8	39.82	84.74	4100	620	3352	128	1085	yes	nonmotile, aerobe
Cylindrospermum stagnale PCC 7417	terrestrial, soil	IV[4]	1 (3)	6229	7.61	42.2	79.71	5659	620	4303	736	1587	yes	photoheterotroph, aerobe
Dactylococcopsis salina PCC 8305	fresh water	I	1 (0)	3337	3.78	42.44	80.37	3129	620	2243	266	998	no	
Escherichia coli O111:H-str. 11128	homo sapiens	I[5]	1 (5)	5732	5.77	50.44	83.97	4953	404	1545	3004	3462	no	pathogenic
Geitlerinema sp. PCC 7407	unknown[6]	III	1 (0)	3815	4.68	58.46	83.59	3544	620	2663	261	1131	yes*	motile

Gloeobacter kilaueensis JS1	terrestrial[7]	I[7]	1 (0)	4507	4.72	60.54	90.33	4225	620	2899	706	1311	no	no thylakoid membrane
***Gloeobacter violaceus* PCC 7421**	terrestrial	I	1 (0)	4430	4.66	62	89.4	4111	620	2933	558	1150	no	no thylakoid membrane
Gloeocapsa sp. PCC 7428	moderate thermophilic, fresh water	I	1 (4)	5011	5.88	43.36	82.6	4478	620	3505	353	1296	no	non-motile, aerobe
Halothece sp. PCC 7418 *(Aphanothece halophytica* 7418)	fresh water	I	1 (0)	3708	4.18	42.92	84.94	3447	620	2607	220	1020	no*	non-motile, anaerobe
Leptolyngbya sp. PCC 7376	cave, terrestrial	III	1 (0)	4228	5.13	43.87	82.67	3942	620	2725	597	1258	no	non-motile, aerobe
Microcoleus sp. PCC 7113	terrestrial, soil	III	1 (8)	6441	7.97	46.21	81.7	5691	620	4184	887	1685	yes	motile, anaerobe
***Microcystis aeruginosa* NIES-843**	fresh water	I	1 (0)	6311	5.84	42.33	81.36	5894	620	3461	1813	2138	no	toxic, bloom forming, non-motile, aerobe
Nodularia spumigena CCY9414	surface, marine	IV	1 (0)	5295	5.46	41.23	80.36	4881	620	3273	988	1664	yes	toxic, bloom forming
Nostoc azollae 0708	fresh water, symbiotic	IV	1 (2)	3651	5.49	38.37	51.43	3469	620	2231	618	1204	yes	motile, aerobe, symbiotic with duckweed
Nostoc punctiforme PCC 73102 (*Nostoc punctiforme* ATCC 29133)	fresh water, soil, symbiotic	IV	1 (5)	6690	9.06	41.35	77.2	5933	620	4728	585	1465	yes	motile, aerobe, symbiotic with Macrozamia
Nostoc sp. PCC 7107	fresh water	IV	1 (0)	5237	6.33	40.36	80.9	4711	620	3787	304	1240	yes	aerobe
***Nostoc* sp. PCC 7120 *(Anabaena* sp. PCC 7120)**	fresh water	IV	1 (6)	6132	7.21	41.27	82.14	5525	620	4437	468	1387	yes	aerobe
Nostoc sp. PCC 7524 *(Nostoc* sp. ATCC 29411)	fresh water	IV	1 (2)	5449	6.72	41.53	81.78	4910	620	3952	338	1318	yes	aerobe, moderate thermal springs
Oscillatoria acuminata PCC 6304	terrestrial, soil	III	1 (2)	5796	7.8	47.61	79.64	5211	620	3858	733	1491	no	aerobe
Oscillatoria nigro-viridis PCC 7112	terrestrial, soil	III	1 (5)	6360	8.27	45.78	77.39	5695	620	4114	961	1687	no	
Pleurocapsa sp. PCC 7327	fresh water, thermophilic	II	1 (0)	4268	4.99	45.19	81.06	3832	620	2861	351	1202	unknown* (yes)	
Prochlorococcus marinus str. AS9601	marine	I	1 (0)	1921	1.67	31.32	90.44	1881	620	1220	41	678	no	50m, HL
***Prochlorococcus marinus* str. MIT 9211**	marine	I	1 (0)	1855	1.69	38.01	89.84	1811	620	1055	136	731	no	83m, LL
***Prochlorococcus marinus* str. MIT 9215**	marine	I	1 (0)	1983	1.74	31.15	89.1	1931	620	1233	78	710	no	5m, HL
Prochlorococcus marinus str. MIT 9301	marine	I	1 (0)	1907	1.64	31.34	90.61	1868	620	1204	44	680	no	90m, HL
Prochlorococcus marinus str. MIT 9303	marine	I	1 (0)	2997	2.68	50.01	84.22	2911	620	1899	392	971	no	135m, HL
Prochlorococcus marinus str. MIT 9312	marine	I	1 (0)	1962	1.71	31.21	90.22	1919	620	1225	74	687	no	135m, LL
Prochlorococcus marinus str. MIT 9313	marine	I	1 (0)	2915	2.41	50.74	85.38	2850	620	1808	422	1015	no	135 m, LL
Prochlorococcus marinus str. MIT 9515	marine	I	1 (0)	1906	1.7	30.79	88.32	1860	620	1181	59	674	no	15m, HL

Prochlorococcus marinus str. NATL1A	marine	I	1 (0)	2193	1.86	34.98	86.6	2138	620	1454	64	692	no	30m, LL
Prochlorococcus marinus str. NATL2A	marine	I	1 (0)	2163	1.84	35.12	87.01	2106	620	1439	47	684	no	30m, LL
Prochlorococcus marinus subsp. marinus str. CCMP1375 (*Prochlorococcus marinus* SS120)	marine	I	1 (0)	1882	1.75	36.44	88.66	1834	620	1093	121	716	no	120m depth, very low light adapted
***Prochlorococcus marinus* subsp. pastoris str. CCMP1986 (*Prochlorococcus marinus* MED4)**	marine	I	1 (0)	1960	1.66	30.8	90.09	1922	620	1207	95	715	no	5m depth, High-light adapted
Prochlorococcus sp. MIT 0604	marine	I[5]	1 (0)	2063	1.78	31.17	87.63	2005	620	1229	156	772	no	135m depth
Prochlorococcus sp. MIT 0801	marine	I[5]	1 (0)	2287	1.93	34.91	85.74	2222	620	1390	212	794	no	40m depth, Low light adapted
Pseudanabaena sp. PCC 7367	marine, intertidal[6]	III	1 (1)	3854	4.89	46.22	77.63	3590	620	2445	525	1170	no	
Rivularia sp. PCC 7116	marine	IV	1 (2)	6644	8.73	37.53	78.16	5900	620	4503	777	1609	unknown* (yes)	motile, heterotroph, aerobe
Stanieria cyanosphaera PCC 7437	fresh water	II	1 (5)	4781	5.54	36.22	81.91	4317	620	3275	422	1239	no	motile
Synechococcus elongatus PCC 6301 (*Synechococcus leopoliensis* SAG 1402-1)	fresh water	I	1 (0)	2525	2.7	55.48	87.65	2406	620	1767	19	829	no	motile, facultative
***Synechococcus elongatus* PCC 7942**	fresh water	I	1 (1)	2661	2.74	55.43	88.81	2537	620	1824	93	888	no	motile, facultative
Synechococcus sp. CC9311	marine, neritic	I	1 (0)	2892	2.61	52.45	86.53	2774	620	1668	486	1144	no	motile
Synechococcus sp. CC9605	marine	I	1 (0)	2638	2.51	59.22	86.57	2542	620	1719	203	887	no	motile
Synechococcus sp. CC9902	marine	I	1 (0)	2304	2.23	54.16	89.63	2218	620	1524	74	786	no	motile
***Synechococcus* sp. JA-2-3B'a(2-13)**	fresh water, thermophilic	I	1 (0)	2862	3.05	58.45	84.97	2743	620	1837	286	989	yes[8]	motile, facultative
Synechococcus sp. JA-3-3Ab	fresh water, thermophilic	I	1 (0)	2760	2.93	60.24	84.47	2628	620	1771	237	918	yes[8]	motile, facultative
Synechococcus sp. KORDI-100	marine, oligotroph	I[5]	1 (0)	3061	2.79	57.5	85.56	2965	620	1788	557	1185	no	aerobe
Synechococcus sp. KORDI-49	marine, mesophile	I[5]	1 (0)	2734	2.59	61.37	87.74	2644	620	1700	324	1002	no	aerobe
Synechococcus sp. KORDI-52	marine	I[5]	1 (0)	2820	2.57	59.09	84.46	2729	620	1769	340	1011	no	aerobe
Synechococcus sp. PCC 6312	fresh water	I	1 (1)	3545	3.72	48.5	84.67	3335	620	2256	459	1162	no	non-motile, aerobe
***Synechococcus* sp. PCC 7002**	mud sample	I	1 (6)	3186	3.41	49.19	87.29	2999	620	2144	235	1028	no	motile, facultative
Synechococcus sp. PCC 7502	fresh water, wetland	I	1 (2)	3318	3.58	40.62	83.6	3123	620	2133	370	1061	no	symbiosis within sphagnum bog

Synechococcus sp. RCC307	marine	I	1 (0)	2535	2.22	60.84	94.53	2443	620	1485	338	996	no	motile, facultative
***Synechococcus* sp. WH 7803**	marine	I	1 (0)	2533	2.37	60.24	93.08	2428	620	1662	146	860	no	motile
Synechococcus sp. WH 8109	marine	I	1 (0)	2644	2.11	60.09	87.91	2572	620	1548	404	1014	no	motile, facultative
***Synechocystis* sp. PCC 6803**	fresh water	I	1 (4)	3564	3.95	47.37	86.64	3314	620	2397	297	1062	no	motile, facultative
***Thermosynechococcus elongatus* BP-1**	thermophilic, fresh water	I	1 (0)	2475	2.59	53.92	89.79	2314	620	1586	108	802	no	
Thermosynechococcus sp. NK55	thermophilic[9]	I[9]	1 (0)	2233	2.52	53.81	85.07	2118	620	1476	22	750	no	
***Trichodesmium erythraeum* IMS101**	marine, neritic	III	1 (0)	4451	7.75	34.14	59.86	4196	620	2808	768	1167	yes	motile, aerobe
cyanobacterium UCYN-A	symbiotic, marine[10]	I	1 (0)	1200	1.44	31.12	80.48	1173	476	664	33	495	yes[10]	symbiosis with *Braarudosphaera bigelowii*; no photosystem II, RuBisCO, or TCA cycle[11]

References:

[1] Hao Wang et al. *Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium Anabaena sp. strain 90.* BMC Genomics 2012, 13:613

[2] Janne Isojärvi et al. *Draft Genome Sequence of Calothrix Strain 336/3, a Novel H2-Producing Cyanobacterium Isolated from a Finnish Lake.* Genome Announcement 2015, vol. 3

[3] Ben de Winder et al. *Crinalium epipsammum sp. nov.: a filamentous cyanobacterium with trichomes composed of elliptical cells and containing poly-p-(1,4) glucan (cellulose).* Microbiology 1990, 136

[4] Joseph Seckbach (Ed.) *Enigmatic Microorganisms and Life in Extreme Environments.* Springer Science, vol. 1

[5] Victor M Markowitz et al. *IMG 4 version of the integrated microbial genomes comparative analysis system.* Nucleic acids research 2013

[6] Rosmarie Rippka et al. *Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria.* Microbiology 1979, 111

[7] Jimmy HW Saw et al. *Cultivation and Complete Genome Sequencing of Gloeobacter kilaueensis sp. nov., from a Lava Cave in K_lauea Caldera, Hawai'i.* 2013, e76376

[8] Anindita Bandyopadhyay et al. *Novel Metabolic Attributes of the Genus Cyanothece, Comprising a Group of Unicellular Nitrogen-Fixing Cyanobacteria.* mBio 2011, e00214

[9] Sergey Stolyar et al. *Genome Sequence of the Thermophilic Cyanobacterium Thermosynechococcus sp. Strain NK55a.* Genome Announcement 2014, vol. 2

[10] Anne W Thompson et al. *Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga.* Science 2012, 337

[11] Kyoko Hagino et al. *Discovery of an Endosymbiotic Nitrogen-Fixing Cyanobacterium UCYN-A in Braarudosphaera bigelowii (Prymnesiophyceae).* PLOSone 2013, e81749

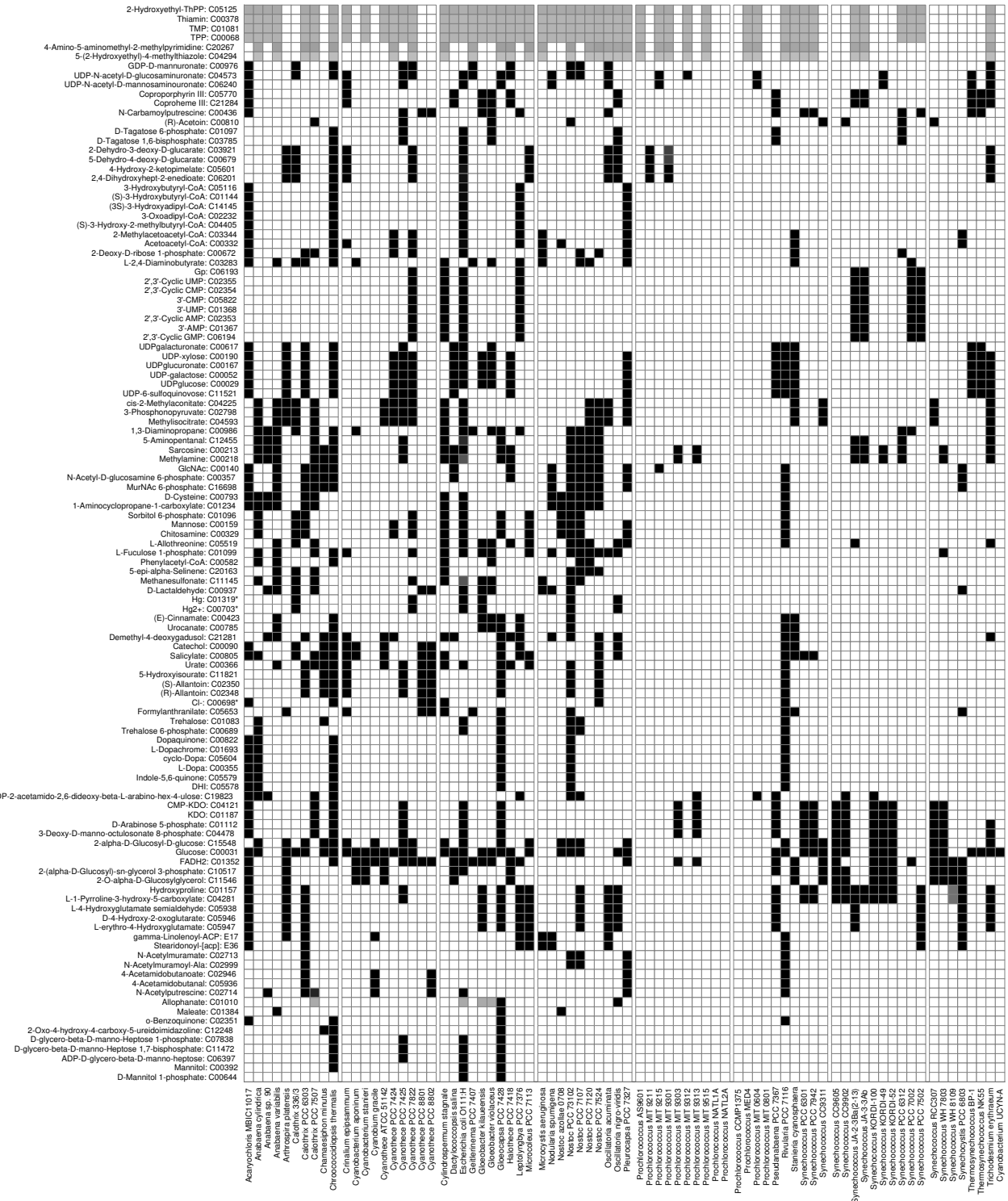
[12] Patrick M Shih et al. *Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing.* PNAS 2013, 110(3)

Appendix B.

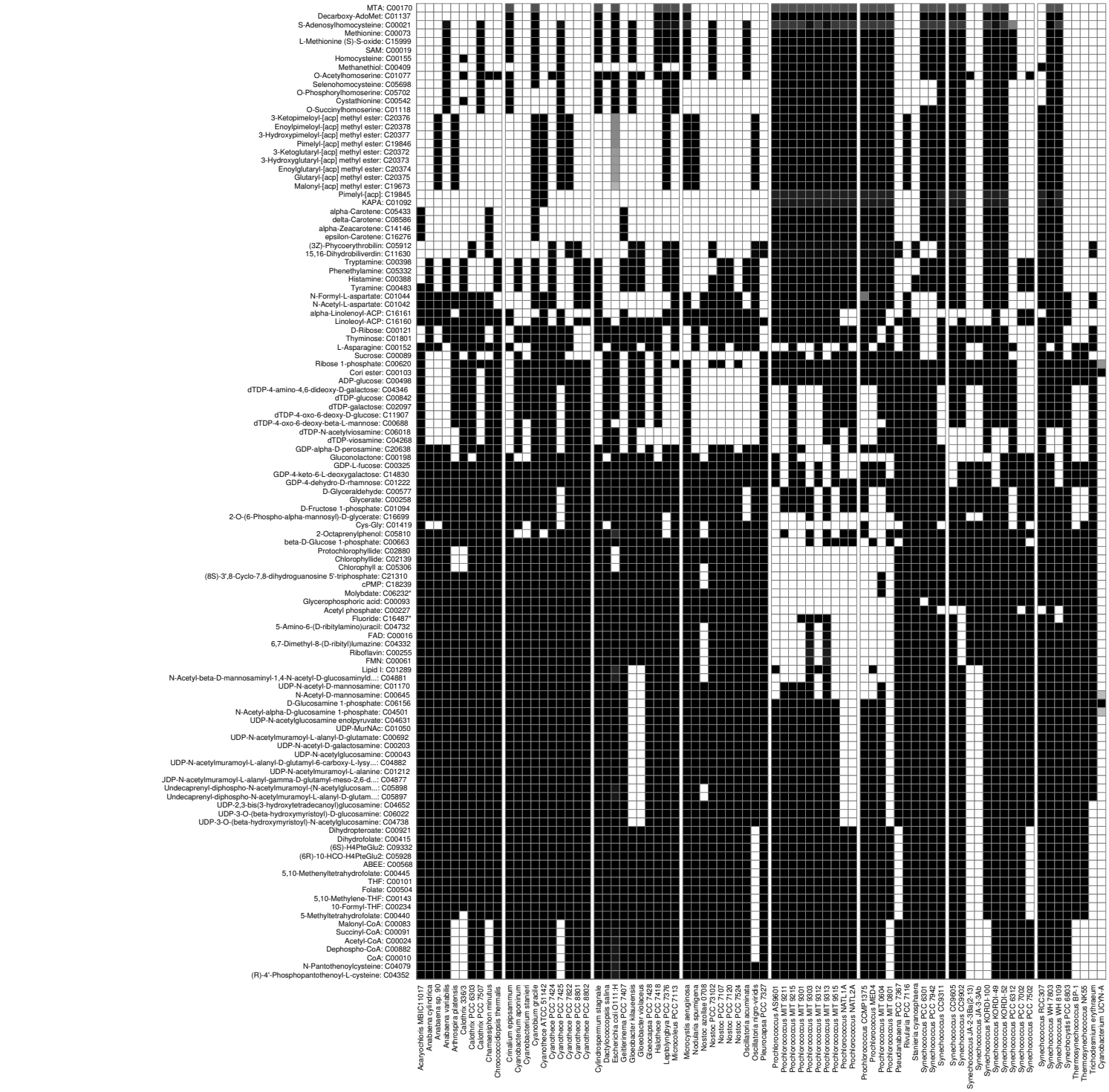
Table of strain-specific synthesis capacities

This table shows the capacity of each strain to synthesize any of the 881 metabolites synthesizeable by the metabolic pan-network. Each box shows the maximal stoichiometric yield for the compound indicated on the y-axis by the species on the x-axis. Yields of carbonaceous compounds are calculated relative to the influx of inorganic carbon. Yields of metabolites containing no carbon were calculated relative to the maximal flux for this compound in simulations of the pan-network. The range of the yield is indicated by the scale on the last page of this table. Boxes are fully black if all available carbon was converted to the according metabolite. White boxes on the other hand, indicate metabolites that could not be synthesized by the according strain, due to the absence of essential enzymes involved in the biosynthesis pathway. Species are sorted alphabetically and denoted at the bottom of the x-axis. Metabolites are sorted according to their profile of synthesizing organisms to help identifying common metabolic pathways. We used the standardized euclidean distance and applied Ward's minimum variance criterion, both implemented in the *linkage* function of MATLAB. All metabolites are denoted by one of their trivial names and the according KEGG identifier on the left y-axis. Metabolites containing no carbon are marked with an asterisk.

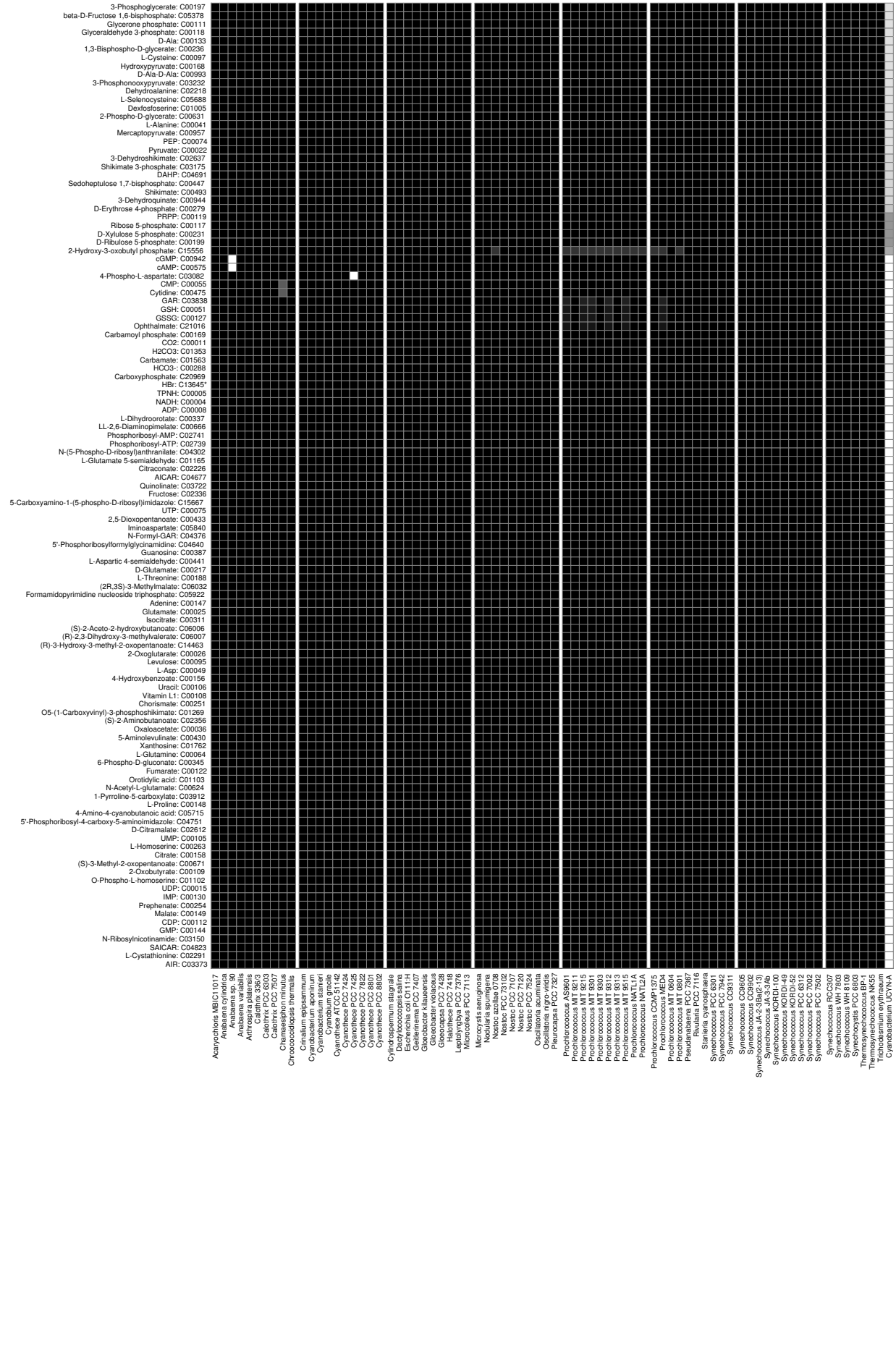

L-Lactate: C00186
3-Mercaptolactate: C05823
2-Hydroxybutyrate: C05984
HFA: C06108
Chloroacetic acid: C06755
2-Chloroethanal: C06754
Bios I: C00137
Inositol 3-phosphate: C04006
L-Threonine phosphate: C12147
Dihydroneopterin phosphate: C05925
Oxalate: C00209
(R)-1-Aminopropan-2-ol: C03194
D-1-Aminopropan-2-ol O-phosphate: C04122
Deoxyinosine: C05512
D-Glutamine: C00819
FMNH2: C01847
Niacin: C00253
2-Hydroxyphenylacetate: C05852
Homogentisate: C00544
delta-Tocotrienol: C14156
delta-Tocopherol: C14151
2-Methyl-6-phytylquinol: C15882
6-Geranylgeranyl-2-methylbenzene-1,4-diol: C20737
2,3-Dimethyl-5-phytylquinol: C15883
beta-Tocotrienol: C14154
gamma-Tocotrienol: C14155
beta-Tocopherol: C14152
6-Geranylgeranyl-2,3-dimethylbenzene-1,4-diol: C20738
alpha-Tocotrienol: C14153
Vitamin E: C02477
gamma-Tocopherol: C02483
Vitamin K1: C02059
Demethylphylloquinone: C13309
1,4-Dihydroxy-2-naphthoate: C03657
Succinylbenzoate: C02730
SHCHC: C05817
1,4-Dihydroxy-2-naphthoyl-CoA: C15547
Succinylbenzoyl-CoA: C03160
5'-Phosphoguanylyl(3'->5')guanosine: C18076
cdiGMP: C16463
S-Formylglutathione: C01031
Methylglyoxal: C00546
L-Lactaldehyde: C00424
S-D-Lactoylglutathione: C03451
D-Lactate: C00256
Fe3+: C14819*
CH3OH: C00132
Bilirubin: C00486
Acetamide: C06244
N-Acetyl-alpha-D-galactosaminyl-diphospho-ditrans,octacis...: C20620
Propanoyl-CoA: C00100
Propanoate: C00163
Propionyladenylate: C05983
Propanoyl phosphate: C02876
Vitamin PP: C00153
Purine-6-ol: C00262
Echinenone: C08592
3-Hydroxyechinenone: C15966
2-Inosose: C00691
D-Aspartate: C00402
D-Gluconate: C00257
Sorbitol: C00794
Xylitol: C00379
Dodecanoate: C02679
Octylic acid: C06423
Decanoate: C01571
Squalene: C00751
Presqualene diphosphate: C03428
22-Hopanol: C06309
Diploptene: C06310
Homocitrate: C01251
e-: C05359*
Diimine: C05360*
Hydrazine: C05361*
N2: C00697*
2-Phenylacetamide: C02505
Indole-3-acetamide: C02693
IAA: C00954
Indoleacetaldehyde: C00637
3-Indoleacetonitrile: C02938
Phenylacetaldehyde: C00601
Phenylacetate: C07086
Benzyl cyanide: C16074
4-Hydroxyphenylacetaldehyde: C03765
4-Hydroxyphenylacetate: C00642
2-Oxoarginine: C03771
4-Guanidinobutanoate: C01035
4-Guanidinobutanamide: C03078
Urea: C00086
CDP-glucose: C00501
CDP-4-dehydro-6-deoxy-D-glucose: C01219
CDP-4-dehydro-3,6-dideoxy-D-glucose: C04297
D-glycero-beta-D-manno-Heptose 7-phosphate: C07836
D-glycero-D-manno-Heptose 7-phosphate: C19882
D-glycero-alpha-D-manno-Heptose 7-phosphate: C19878
D-glycero-alpha-D-manno-Heptose 1-phosphate: C19880
D-glycero-alpha-D-manno-Heptose 1,7-bisphosphate: C19879
GDP-D-glycero-alpha-D-manno-heptose: C19881
alpha-D-Galactose 1-phosphate: C00446
NH2OH: C00192*
D-2,3-Diketo-4-deoxy-epi-inositol: C04287
D-Psicose: C06468
Phosphite: C06701*
Dimethylbenzimidazole: C03114
alpha-Ribazole: C05775
alpha-Ribazole 5'-phosphate: C04778
(Sulfide)n: C19692*
7,8-Didemethyl-8-hydroxy-5-deazariboflavin: C19154
4-Amino-2-methyl-5-(phosphooxymethyl)pyrimidine: C04556
4-Cresol: C01468
4-Methyl-5-(2-phosphooxyethyl)thiazole: C04327
Glycolaldehyde: C00266
4-Amino-2-methyl-5-(diphosphooxymethyl)pyrimidine: C04752
Toxopyrimidine: C01279
2-(1,2-Dihydroxyethyl)-TPP: C13378
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Cyanothece PA 3822
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cylindrospermum stagnale
Cyanothece ATCC 51142
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Escherichia coli O111:H
Geitlerinema PCC 7407
Gloeobacter kilaueensis
Gloeobacter violaceus
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya PCC 7376
Microcoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 73102
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc PCC 7524
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 6301
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-100
Synechococcus KORDI-34
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus PCC 7502
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechocystis PCC 6803
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A

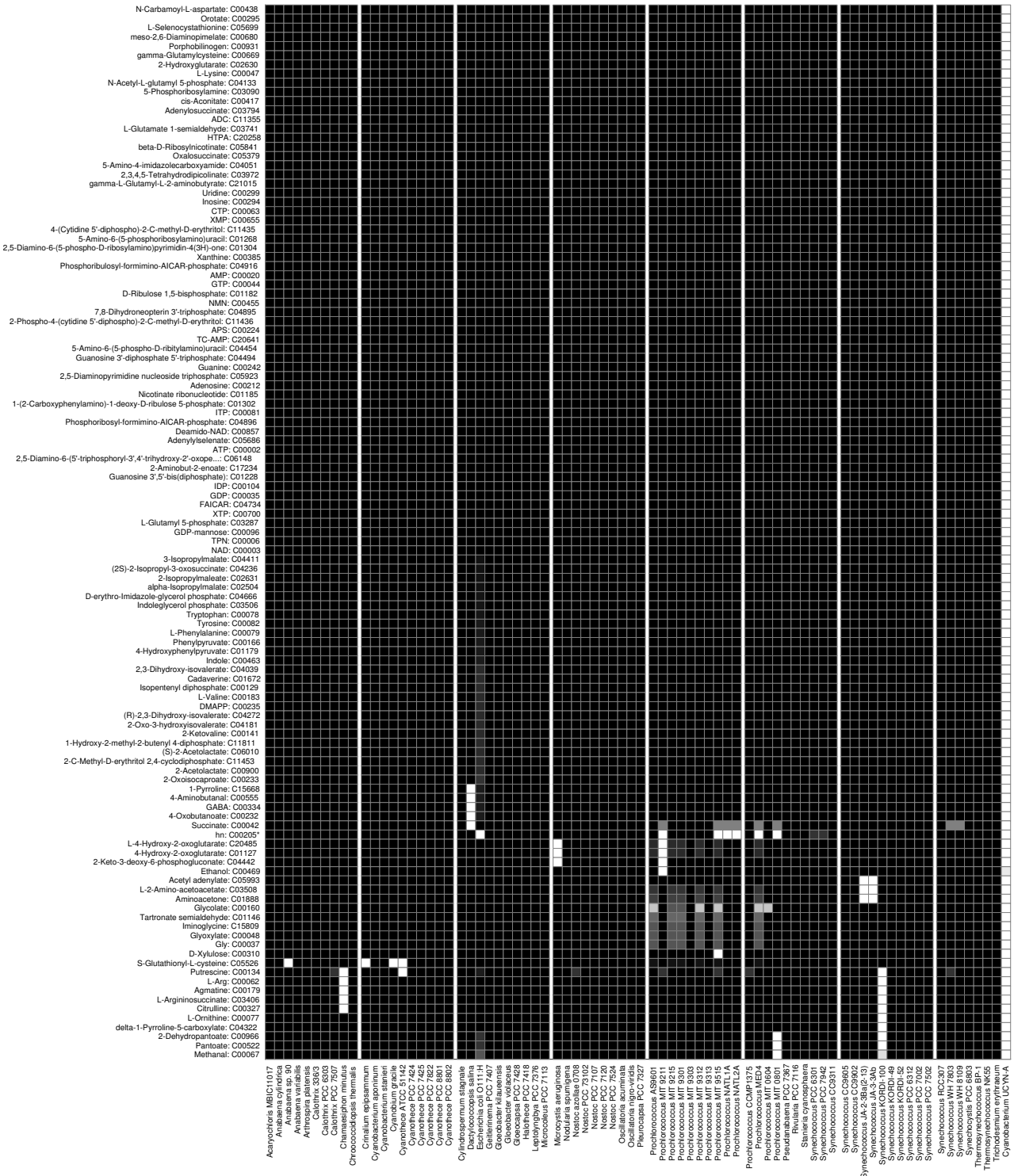

2-Hydroxyethyl-ThPP: C05125
Thiamin: C00378
TMP: C01081
TPP: C00068
4-Amino-5-aminomethyl-2-methylpyrimidine: C00267
5-(2-Hydroxyethyl)-4-methylthiazole: C04294
GDP-D-mannuronate: C00976
UDP-N-acetyl-D-glucosaminuronate: C04573
UDP-N-acetyl-D-mannosaminouronate: C06240
Coproporphyrin III: C05770
Coproheme III: C21284
N-Carbamoylputrescine: C00436
(R)-Acetoin: C00810
D-Tagatose 6-phosphate: C01097
D-Tagatose 1,6-bisphosphate: C03785
2-Dehydro-3-deoxy-D-glucarate: C03921
5-Dehydro-4-deoxy-D-glucarate: C00679
4-Hydroxy-2-ketopimelate: C05601
2,4-Dihydroxyhept-2-enedioate: C06201
3-Hydroxybutyryl-CoA: C05116
(S)-3-Hydroxybutyryl-CoA: C01144
(3S)-3-Hydroxyadipyl-CoA: C14145
3-Oxoadipyl-CoA: C02232
(S)-3-Hydroxy-2-methylbutyryl-CoA: C04405
2-Methylacetoacetyl-CoA: C03344
Acetoacetyl-CoA: C00332
2-Deoxy-D-ribose 1-phosphate: C00672
L-2,4-Diaminobutyrate: C03283
Gp: C06193
2',3'-Cyclic UMP: C02355
2',3'-Cyclic CMP: C02354
3'-CMP: C05822
3'-UMP: C01368
2',3'-Cyclic AMP: C02353
3'-AMP: C01367
2',3'-Cyclic GMP: C06194
UDPgalacturonate: C00617
UDP-xylose: C00190
UDPglucuronate: C00167
UDP-galactose: C00052
UDPglucose: C00029
UDP-6-sulfoquinovose: C11521
cis-2-Methylaconitate: C04225
3-Phosphonopyruvate: C02798
Methylisocitrate: C04593
1,3-Diaminopropane: C00986
5-Aminopentanal: C12455
Sarcosine: C00213
Methylamine: C00218
GlcNAc: C00140
N-Acetyl-D-glucosamine 6-phosphate: C00357
MurNAc 6-phosphate: C16698
D-Cysteine: C00793
1-Aminocyclopropane-1-carboxylate: C01234
Sorbitol 6-phosphate: C01096
Mannose: C00159
Chitosamine: C00329
L-Allothreonine: C05519
L-Fuculose 1-phosphate: C01099
Phenylacetyl-CoA: C00582
5-epi-alpha-Selinene: C20163
Methanesulfonate: C11145
D-Lactaldehyde: C00937
Hg: C01319*
Hg2+: C00703*
(E)-Cinnamate: C00423
Urocanate: C00785
Demethyl-4-deoxygadusol: C21281
Catechol: C00090
Salicylate: C00805
Urate: C00366
5-Hydroxyisourate: C11821
(S)-Allantoin: C02350
(R)-Allantoin: C02348
Cl-: C06898*
Formylanthranilate: C05653
Trehalose: C01083
Trehalose 6-phosphate: C00689
Dopaquinone: C00822
L-Dopachrome: C01693
cyclo-Dopa: C05604
L-Dopa: C00355
Indole-5,6-quinone: C05579
DHI: C05578
UDP-2-acetamido-2,6-dideoxy-beta-L-arabino-hex-4-ulose: C19823
CMP-KDO: C04121
KDO: C01187
D-Arabinose 5-phosphate: C01112
3-Deoxy-D-manno-octulosonate 8-phosphate: C04478
2-alpha-D-Glucosyl-D-glucose: C15548
Glucose: C00031
FADH2: C01352
2-(alpha-D-Glucosyl)-sn-glycerol 3-phosphate: C10517
2-O-alpha-D-Glucosylglycerol: C11546
Hydroxyproline: C01157
L-1-Pyrroline-3-hydroxy-5-carboxylate: C04281
L-4-Hydroxyglutamate semialdehyde: C05938
D-4-Hydroxy-2-oxoglutarate: C05946
L-erythro-4-Hydroxyglutamate: C05947
gamma-Linolenoyl-ACP: E17
Stearidonoyl-[acp]: E36
N-Acetylmuramate: C02713
N-Acetylmuramoyl-Ala: C02999
4-Acetamidobutanoate: C02946
4-Acetamidobutanal: C05936
N-Acetylputrescine: C02714
Allophanate: C01010
Maleate: C01384
o-Benzoquinone: C02351
2-Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline: C12248
D-glycero-beta-D-manno-Heptose 1-phosphate: C07838
D-glycero-beta-D-manno-Heptose 1,7-bisphosphate: C11472
ADP-D-glycero-beta-D-manno-heptose: C06397
Mannitol: C00392
D-Mannitol 1-phosphate: C00644
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Calothrix 336/3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanobium gracile
Oxynema PCC 7342
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Fischerella PCC 9431
Geitleriema PCC 7407
Gloeobacter kilaueensis
Gloeobacter violaceus
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya PCC 7376
Microcoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 73102
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc PCC 7524
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 9604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 6301
Synechococcus PCC 7002
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-100
Synechococcus KORDI-49
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus PCC 7502
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechocystis PCC 6803
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A

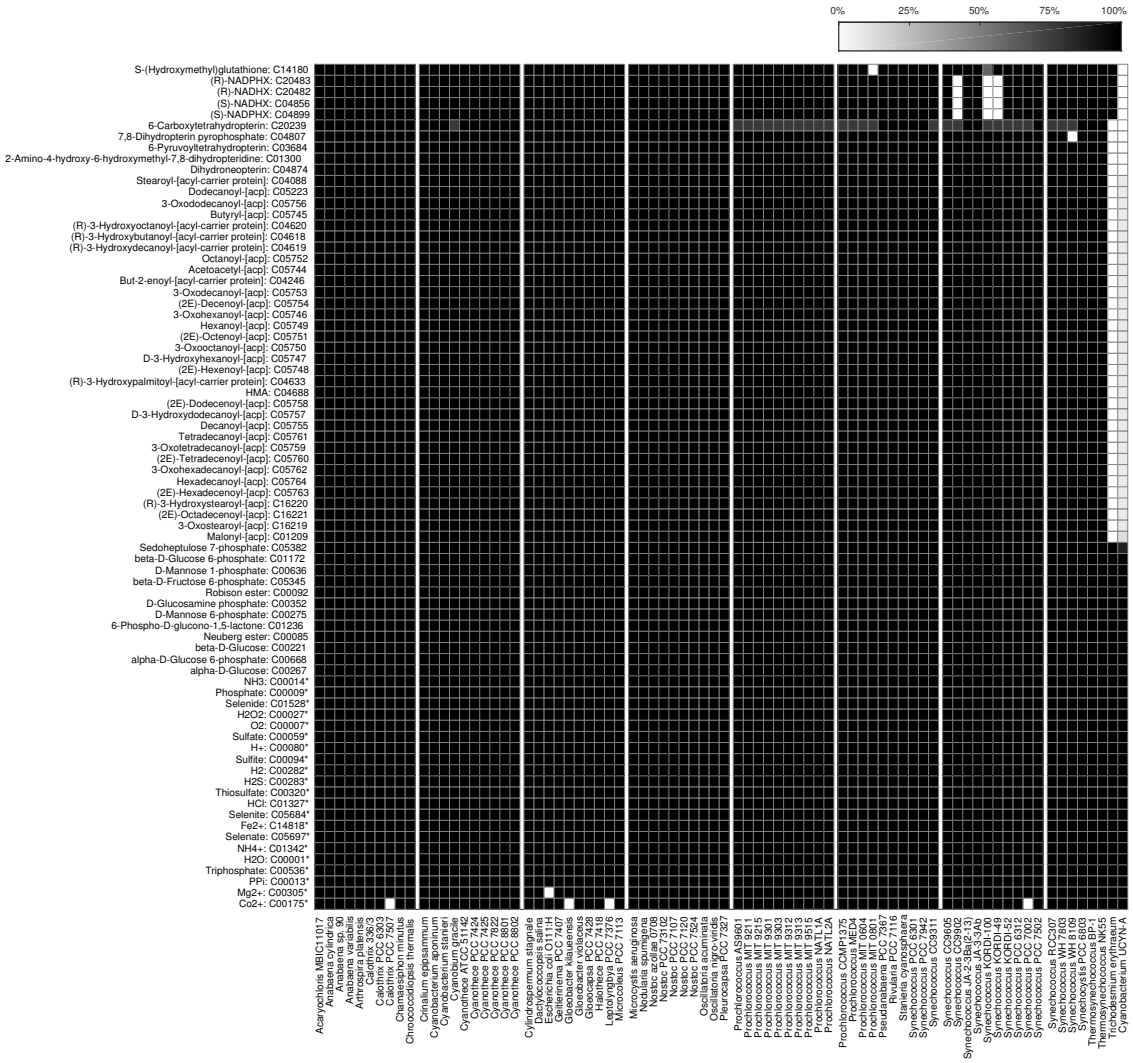

ADP-L-glycero-beta-D-manno-heptose: C06398
Ribose 1,5-bisphosphate: C01151
(2S,3S)-2,3-Dihydro-2,3-dihydroxybenzoate: C04171
CDP-glycerol: C00513
4-Hydroxy-2-oxovalerate: C03589
Maltose: C00208
Maltose 6'-phosphate: C02995
Acetyl-maltose: C02130
Amylose: C00718
6-Phosphosucrose: C16688
(-)-Ureidoglycolate: C00603
Glycocyamine: C00581
L-beta-Lysine: C01142
Nitrite: C00088*
N6-Hydroxy-L-lysine: C01028
Methylselenol: C05703
Se-Methyl-L-selenocysteine: C05689
L-Selenomethionine: C05335
Se-Adenosylselenomethionine: C05691
7,8-Dihydromonapterin 3'-triphosphate: C21094
L-Methionine (R)-S-oxide: C15998
L-Methionine S-oxide: C02989
(4Z)-Hexadec-4-enoyl-[acyl-carrier protein]: C21114
Sapienoyl-[acyl-carrier-protein]: C21115
Allantoate: C00499
Phenol: C00146
Aryl sulfate: C00850
Dehydroascorbate: C05422
L-xylo-Hexulonolactone: C03289
L-Ascorbate 6-phosphate: C16186
D-Fructuronate: C00905
L-Gulonolactone: C01040
Glucuronate: C00191
Ascorbate radical: C01041
Ascorbate: C00072
2,3-Diketo-L-gulonate: C04575
Gulonate: C00800
L-Kynurenine: C00328
Geosmin: C16286
Germacradienol: C16143
(-)-Germacrene D: C16142
Dopamine: C03758
Protocatechuatealdehyde: C04043
Homoprotocatechuate: C01161
L-Formylkynurenine: C02700
UDP-3-keto-GlcNAcA: C20395
UDP-D-galacto-1,4-furanose: C03733
4-Maleylacetoacetate: C01036
2-O-(alpha-D-Glucopyranosyl)-3-phospho-D-glycerate: C19791
N-Acetylmannosamine 6-phosphate: C04257
Zeaxanthin diglucoside: C15969
(R)-3-Hydroxybutanoyl-CoA: C03561
Oxidized plastocyanin: C03162*
Reduced plastocyanin: C03025*
7-Chloro-L-tryptophan: C19687
Dialurate: E20
Saccharopine: C00449
Methyl sulfide: C00580
Acetone: C00207
Taurocyamine: C01959
Choline sulfate: C00919
HMG-CoA: C00356
3-Hydroxyphenylacetate: C05593
Cinnavalininate: C05640
NAAG: C12270
NAAG2: C20776
D-Glucaro-1,5-lactone: C20890
2-Sulfotrehalose: C20985
Sulfoquinovose: C20829
S-Methyl-5'-thioinosine: C19787
O2-: C00704*
S-Inosyl-L-homocysteine: C03431
3-Methylglutaconyl-CoA: C03231
3-Hydroxy-L-kynurenine: C03227
3-Hydroxyanthranilate: C00632
Betaine aldehyde: C00576
Choline: C00114
4-Hydroxybenzaldehyde: C00633
CO: C00237
Ethylnitronate: C18091
2-Hydroxypropyl-ThPP: C21017
2-Methyl-1-hydroxypropyl-TPP: C15976
3-Methyl-1-hydroxybutyl-TPP: C15974
2-Methyl-1-hydroxybutyl-TPP: C15978
Dehydrospermidine: C15853
Homospermidine: C06366
Glutathionylspermidine: C05730
4-Imidazolone-5-propanoate: C03680
N-Formimino-L-glutamate: C00439
5-Formiminotetrahydrofolate: C00664
Sucrose 6'-phosphate: C02591
3-[(4R)-4-Hydroxycyclohexa-1,5-dien-1-yl]-2-oxopropanoate: C20953
3-[(1E,4R)-4-Hydroxycyclohex-2-en-1-ylidene]-2-oxopropanoate: C21085
3-Sulfopyruvate: C05528
L-Cysteate: C00506
Taurine: C00245
(2R)-Phosphosulfolactate: C11536
(R)-3-Sulfolactate: C11537
5-Methylthio-D-ribose: C03089
S-Ribosylhomocysteine: C03539
Formamide: C00488
Dimethylglycine: C01026
Betaine: C00719
3-Methylthiopropionate: C08276
4-Methylthio-2-oxobutanoate: C01180
2-Hydroxy-3-keto-5-methylthiopentenyl-1-phosphate: C15651
1,2-Dihydroxy-5-(methylthio)pent-1-en-3-one: C15606
5-(Methylthio)-2,3-dioxopentyl phosphate: C15650
S-Methyl-5-thio-D-ribulose 1-phosphate: C04582
DAPA: C01037
S-Adenosyl-4-methylthio-2-oxobutanoate: C04425
S-Methyl-5-thio-D-ribose 1-phosphate: C04188
Spermine: C00750
Spermidine: C00315
Aminopropylcadaverine: C16565
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Calothrix 336/3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanobium gracile
Cyanothece ATCC 51142
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Escherichia coli O111:H
Gloeobacter kilaueensis
Gloeobacter violaceus
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya PCC 7376
Mastigocoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 73102
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc PCC 7524
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus COMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 6301
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-100
Synechococcus KORDI-49
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechocystis PCC 6803
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A


MTA: C00170
Decarboxy-AdoMet: C01137
S-Adenosylhomocysteine: C00021
Methionine: C00073
L-Methionine (S)-S-oxide: C15999
SAM: C00019
Homocysteine: C00155
Methanethiol: C00409
O-Acetylhomoserine: C01077
Selenohomocysteine: C05698
O-Phosphorylhomoserine: C05702
Cystathionine: C00542
O-Succinylhomoserine: C01118
3-Ketopimeloyl-[acp] methyl ester: C20376
Enoylpimeloyl-[acp] methyl ester: C20378
3-Hydroxypimeloyl-[acp] methyl ester: C20377
Pimelyl-[acp] methyl ester: C19846
3-Ketoglutaryl-[acp] methyl ester: C20372
3-Hydroxyglutaryl-[acp] methyl ester: C20373
Enoylglutaryl-[acp] methyl ester: C20374
Glutaryl-[acp] methyl ester: C20375
Malonyl-[acp] methyl ester: C19673
Pimelyl-[acp]: C19845
KAPA: C01092
alpha-Carotene: C05433
delta-Carotene: C08596
alpha-Zeacarotene: C14146
epsilon-Carotene: C16276
(3Z)-Phycoerythrobilin: C05912
15,16-Dihydrobiliverdin: C11630
Tryptamine: C00398
Phenethylamine: C05332
Histamine: C00388
Tyramine: C00483
N-Formyl-L-aspartate: C01044
N-Acetyl-L-aspartate: C01042
alpha-Linolenoyl-ACP: C16161
Linoleoyl-ACP: C16160
D-Ribose: C00121
Thyminose: C01801
L-Asparagine: C00152
Sucrose: C00089
Ribose 1-phosphate: C00620
Cori ester: C00103
ADP-glucose: C00498
dTDP-4-amino-4,6-dideoxy-D-galactose: C04346
dTDP-glucose: C00842
dTDP-galactose: C02097
dTDP-4-oxo-6-deoxy-D-glucose: C11907
dTDP-4-oxo-6-deoxy-beta-L-mannose: C00688
dTDP-N-acetylviosamine: C06018
dTDP-viosamine: C04268
GDP-alpha-D-perosamine: C20638
Gluconolactone: C00198
GDP-L-fucose: C00325
GDP-4-keto-6-L-deoxygalactose: C14830
GDP-4-dehydro-D-rhamnose: C01222
D-Glyceraldehyde: C00577
Glycerate: C00258
D-Fructose 1-phosphate: C01094
2-O-(6-Phospho-alpha-mannosyl)-D-glycerate: C16699
Cys-Gly: C01419
2-Octaprenylphenol: C05810
beta-D-Glucose 1-phosphate: C00663
Protochlorophyllide: C02880
Chlorophyllide: C02139
Chlorophyll a: C05306
(8S)-3',8-Cyclo-7,8-dihydroguanosine 5'-triphosphate: C21310
cPMP: C18239
Molybdate: C06232*
Glycerophosphoric acid: C00093
Acetyl phosphate: C00227
Fluoride: C16487*
5-Amino-6-(D-ribitylamino)uracil: C04732
FAD: C00016
6,7-Dimethyl-8-(D-ribityl)lumazine: C04332
Riboflavin: C00255
FMN: C00061
Lipid I: C01289
N-Acetyl-beta-D-mannosaminyl-1,4-N-acetyl-D-glucosaminyld...: C04881
UDP-N-acetyl-D-mannosamine: C01170
N-Acetyl-D-mannosamine: C00645
D-Glucosamine 1-phosphate: C06156
N-Acetyl-alpha-D-glucosamine 1-phosphate: C04501
UDP-N-acetylglucosamine enolpyruvate: C04631
UDP-MurNAc: C01050
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate: C00692
UDP-N-acetyl-D-galactosamine: C00203
UDP-N-acetylglucosamine: C00043
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-6-carboxy-L-lysy...: C04882
UDP-N-acetylmuramoyl-L-alanine: C01212
JDP-N-acetylmuramoyl-L-alanyl-gamma-D-glutamyl-meso-2,6-d...: C04877
Undecaprenyl-diphospho-N-acetylmuramoyl-(N-acetylglucosam...: C05898
Undecaprenyl-diphospho-N-acetylmuramoyl-L-alanyl-D-glutam...: C05897
UDP-2,3-bis(3-hydroxytetradecanoyl)glucosamine: C04652
UDP-3-O-(beta-hydroxymyristoyl)-D-glucosamine: C06022
UDP-3-O-(beta-hydroxymyristoyl)-N-acetylglucosamine: C04738
Dihydropteroate: C00921
Dihydrofolate: C00415
(6S)-H4PteGlu2: C09332
(6R)-10-HCO-H4PteGlu2: C05928
ABEE: C00568
5,10-Methenyltetrahydrofolate: C00445
THF: C00101
Folate: C00504
5,10-Methylene-THF: C00143
10-Formyl-THF: C00234
5-Methyltetrahydrofolate: C00440
Malonyl-CoA: C00083
Succinyl-CoA: C00091
Acetyl-CoA: C00024
Dephospho-CoA: C00882
CoA: C00010
N-Pantothenoylcysteine: C04079
(R)-4'-Phosphopantothenoyl-L-cysteine: C04352
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Calothrix 336/3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanobium gracile
Cyanothece ATCC 51142
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Escherichia coli O111:H
Geitlerinema PCC 7407
Gloeobacter kilaueensis
Gloeobacter violaceus
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya PCC 7376
Microcoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 73102
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc PCC 7524
Nostoc sp. PCC 7936
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 6301
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-100
Synechococcus KORDI-49
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus PCC 7502
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechocystis PCC 6803
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A

D-4'-Phosphopantothenate: C03492
Phosphopantetheine: C01134
Pantetheine: C00831
Pantothenate: C00864
3-Oxopropanoate: C00222
beta-Alanine: C00099
3-Aminopropanal: C05665
dTDP: C00363
TTP: C00459
Thymidine: C00214
dTMP: C00364
dATP: C00131
dGDP: C00361
dGMP: C00362
Deoxyguanosine: C00330
dAMP: C00360
dGTP: C00286
dUMP: C00365
dCDP: C00705
dUDP: C01346
dUTP: C00460
Deoxyuridine: C00526
dCTP: C00458
Deoxyadenosine: C00559
dCMP: C00239
Deoxycytidine: C00881
dADP: C00206
L-Leucine: C00123
L-Isoleucine: C00407
PAPS: C00053
PAP: C00054
3'-Phosphoadenylylselenate: C05696
5'-Deoxyadenosine: C05198
Selenophosphate: C05172*
AppppA: C01260
9,9'-dicis-zeta-Carotene: C15857
7,9,9'-tricis-Neurosporene: C19759
7,9,7',9'-tetracis-Lycopene: C15858
2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carbox...: C16519
Isochorismate: C00885
9,15,9'-tricis-zeta-Carotene: C19764
15,9'-dicis-Phytofluene: C19765
beta-Cryptoxanthin: C08591
Zeaxanthin: C06098
3-Octaprenyl-4-hydroxybenzoate: C05809
Nonaprenyl-4-hydroxybenzoate: C03885
Undecaprenyl phosphate: C17556
Farnesyl diphosphate: C00448
Bactoprenyl diphosphate: C04574
Geranyl diphosphate: C00341
Geranylgeranyl diphosphate: C00353
Solanesyl diphosphate: C04145
Farnesylfarnesylgeraniol: C04146
7,8-Dihydro-beta-carotene: C16291
Prephytoene diphosphate: C03427
Phytyl diphosphate: C05427
all-trans-zeta-Carotene: C05430
all-trans-Neurosporene: C05431
beta-Zeacarotene: C05434
gamma-Carotene: C05435
Lycopene: C05432
Phytoene: C05421
beta-Carotene: C02094
all-trans-Phytoene: C05413
all-trans-Phytofluene: C05414
2-Phosphoglycolate: C00988
N-Acetylornithine: C00437
2-Acetamido-5-oxopentanoate: C01250
Hexadecenoyl-[acyl-carrier protein]: C16520
Octadecenoyl-[acp]: C01203
7,8-Dihydromonapterin: C21008
Aminoacetaldehyde: C06735
2-Aminomalonate semialdehyde: C11822
Glycerone: C00184
Glycerol: C00116
MSBQ: C17570
Plastoquinone-9: C10385
Plastoquinol-9: C16695
Precorrin 3A: C05772
(3Z)-Phycocyanobilin: C05786
Biliverdin: C00500
Mg-protoporphyrin IX: C03516
Mg-Protoporphyrin IX 13-monomethyl ester: C04536
Divinyl chlorophyllide a: C11832
Divinylprotochlorophyllide: C11831
13(1)-Hydroxy-Mg-protoporphyrin IX 13-monomethyl ester: C11829
13(1)-Oxo-Mg-protoporphyrin IX 13-monomethyl ester: C11830
Heme A: C15670
Heme O: C15672
Precorrin 2: C02463
Precorrin 1: C15527
Protoporphyrinogen IX: C01079
Protoporphyrin: C02191
Heme: C00032
Uroporphyrinogen III: C01051
Coproporphyrinogen III: C03263
Coproporphyrinogen I: C05768
Uroporphyrinogen I: C05766
Hydroxymethylbilane: C01024
L-Histidine: C00135
L-Histidinol phosphate: C01100
L-Histidinal: C01929
L-Histidinol: C00860
Imidazole-acetol phosphate: C01267
Formate: C00058
2-Deoxy-D-ribose 5-phosphate: C00673
1-Deoxy-D-xylulose 5-phosphate: C11437
2-C-Methyl-D-erythritol 4-phosphate: C11434
Acetyl-[acyl-carrier protein]: C03939
Ethanal: C00084
Acetate: C00033
O-Acetyl-L-serine: C00979
Serine: C00065
S-Sulfo-L-cysteine: C05824
D-Fructose 1,6-bisphosphate: C00354
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Calothrix 336/3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanobium gracile
Cyanothece PCC 7424
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Fischerella sp. JSC-11
Geitleriinema PCC 7407
Gloeobacter kilauensis
Gloeobacter violaceus
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya sp. PCC 7376
Microcoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 73102
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc PCC 7524
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 6312
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-100
Synechococcus KORDI-49
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus PCC 7502
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechococcus WH 8102
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A


3-Phosphoglycerate: C00197
beta-D-Fructose 1,6-bisphosphate: C05378
Glycerone phosphate: C00111
Glyceraldehyde 3-phosphate: C00118
D-Ala: C00133
1,3-Bisphospho-D-glycerate: C00236
L-Cysteine: C00097
Hydroxypyruvate: C00168
D-Ala-D-Ala: C00993
3-Phosphonooxypyruvate: C03232
Dehydroalanine: C02218
L-Selenocysteine: C05688
Dexfosfoserine: C01005
2-Phospho-D-glycerate: C00631
L-Alanine: C00041
Mercaptopyruvate: C00957
PEP: C00074
Pyruvate: C00022
3-Dehydroshikimate: C02637
Shikimate 3-phosphate: C03175
DAHP: C04691
Sedoheptulose 1,7-bisphosphate: C00447
Shikimate: C00493
3-Dehydroquinate: C00944
D-Erythrose 4-phosphate: C00279
PRPP: C00119
Ribose 5-phosphate: C00117
D-Xylulose 5-phosphate: C00231
D-Ribulose 5-phosphate: C00199
2-Hydroxy-3-oxobutyl phosphate: C15556
cGMP: C00942
cAMP: C00575
4-Phospho-L-aspartate: C03082
CMP: C00055
Cytidine: C00475
GAR: C03838
GSH: C00051
GSSG: C00127
Ophthalmate: C21016
Carbamoyl phosphate: C00169
CO2: C00011
H2CO3: C01353
Carbamate: C01563
HCO3-: C00288
Carboxyphosphate: C20969
HBr: C13645*
TPNH: C00005
NADH: C00004
ADP: C00008
L-Dihydroorotate: C00337
LL-2,6-Diaminopimelate: C00666
Phosphoribosyl-AMP: C02741
Phosphoribosyl-ATP: C02739
N-(5-Phospho-D-ribosyl)anthranilate: C04302
L-Glutamate 5-semialdehyde: C01165
Citraconate: C02226
AICAR: C04677
Quinolinate: C03722
Fructose: C02336
5-Carboxyamino-1-(5-phospho-D-ribosyl)imidazole: C15667
UTP: C00075
2,5-Dioxopentanoate: C00433
Iminoaspartate: C05840
N-Formyl-GAR: C04376
5'-Phosphoribosylformylglycinamidine: C04640
Guanosine: C00387
L-Aspartic 4-semialdehyde: C00441
D-Glutamate: C00217
L-Threonine: C00188
(2R,3S)-3-Methylmalate: C06032
Formamidopyrimidine nucleoside triphosphate: C05922
Adenine: C00147
Glutamate: C00025
Isocitrate: C00311
(S)-2-Aceto-2-hydroxybutanoate: C06006
(R)-2,3-Dihydroxy-3-methylvalerate: C06007
(R)-3-Hydroxy-3-methyl-2-oxopentanoate: C14463
2-Oxoglutarate: C00026
Levulose: C00095
L-Asp: C00049
4-Hydroxybenzoate: C00156
Uracil: C00106
Vitamin L1: C00108
Chorismate: C00251
O5-(1-Carboxyvinyl)-3-phosphoshikimate: C01269
(S)-2-Aminobutanoate: C02356
Oxaloacetate: C00036
5-Aminolevulinate: C00430
Xanthosine: C01762
L-Glutamine: C00064
6-Phospho-D-gluconate: C00345
Fumarate: C00122
Orotidylic acid: C01103
N-Acetyl-L-glutamate: C00624
1-Pyrroline-5-carboxylate: C03912
L-Proline: C00148
4-Amino-4-cyanobutanoic acid: C05715
5'-Phosphoribosyl-4-carboxy-5-aminoimidazole: C04751
D-Citramalate: C02612
UMP: C00105
L-Homoserine: C00263
Citrate: C00158
(S)-3-Methyl-2-oxopentanoate: C00671
2-Oxobutyrate: C00109
O-Phospho-L-homoserine: C01102
UDP: C00015
IMP: C00130
Prephenate: C00254
Malate: C00149
CDP: C00112
GMP: C00144
N-Ribosylnicotinamide: C03150
SAICAR: C04823
L-Cystathionine: C02291
AIR: C03373
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Calothrix sp. 336/3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanothece ATCC 51142
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Escherichia coli O111:H
Geitlerinema PCC 7407
Gloeobacter kilauensis
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya PCC 7376
Microcoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 7310Z
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc PCC 7524
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Synechococcus elongatus
Synechocystis PCC 6301
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus PCC 7335
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechocystis PCC 6803
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A


N-Carbamoyl-L-aspartate: C00438
Orotate: C00295
L-Selenocystathionine: C05699
meso-2,6-Diaminopimelate: C00680
Porphobilinogen: C00931
gamma-Glutamylcysteine: C00669
2-Hydroxyglutarate: C02630
L-Lysine: C00047
N-Acetyl-L-glutamyl 5-phosphate: C04133
5-Phosphoribosylamine: C03090
cis-Aconitate: C00417
Adenylosuccinate: C03794
ADC: C11355
L-Glutamate 1-semialdehyde: C03741
HTPA: C20258
beta-D-Ribosylnicotinate: C05841
Oxalosuccinate: C05379
5-Amino-4-imidazolecarboxyamide: C04051
2,3,4,5-Tetrahydrodipicolinate: C03972
gamma-L-Glutamyl-L-2-aminobutyrate: C21015
Uridine: C00299
Inosine: C00294
CTP: C00063
XMP: C00655
4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol: C11435
5-Amino-6-(5-phosphoribosylamino)uracil: C01268
2,5-Diamino-6-(5-phospho-D-ribosylamino)pyrimidin-4(3H)-one: C01304
Xanthine: C00385
Phosphoribulosyl-formimino-AICAR-phosphate: C04916
AMP: C00020
GTP: C00044
D-Ribulose 1,5-bisphosphate: C01182
NMN: C00455
7,8-Dihydroneopterin 3'-triphosphate: C04895
2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol: C11436
APS: C00224
TC-AMP: C20641
5-Amino-6-(5-phospho-D-ribitylamino)uracil: C04454
Guanosine 3'-diphosphate 5'-triphosphate: C04494
Guanine: C00242
2,5-Diaminopyrimidine nucleoside triphosphate: C05923
Adenosine: C00212
Nicotinate ribonucleotide: C01185
1-(2-Carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate: C01302
ITP: C00081
Phosphoribosyl-formimino-AICAR-phosphate: C04896
Deamido-NAD: C00857
Adenylylselenate: C05686
ATP: C00002
2,5-Diamino-6-(5'-triphosphoryl-3',4'-trihydroxy-2'-oxope...: C06148
2-Aminobut-2-enoate: C17234
Guanosine 3',5'-bis(diphosphate): C01228
IDP: C00104
GDP: C00035
FAICAR: C04734
XTP: C00700
L-Glutamyl 5-phosphate: C03287
GDP-mannose: C00096
TPN: C00006
NAD: C00003
3-Isopropylmalate: C04411
(2S)-2-Isopropyl-3-oxosuccinate: C04236
2-Isopropylmaleate: C02631
alpha-Isopropylmalate: C02504
D-erythro-Imidazole-glycerol phosphate: C04666
Indoleglycerol phosphate: C03506
Tryptophan: C00078
Tyrosine: C00082
L-Phenylalanine: C00079
Phenylpyruvate: C00166
4-Hydroxyphenylpyruvate: C01179
Indole: C00463
2,3-Dihydroxy-isovalerate: C04039
Cadaverine: C01672
Isopentenyl diphosphate: C00129
L-Valine: C00183
DMAPP: C00235
(R)-2,3-Dihydroxy-isovalerate: C04272
2-Oxo-3-hydroxyisovalerate: C04181
2-Ketovaline: C00141
1-Hydroxy-2-methyl-2-butenyl 4-diphosphate: C11811
(S)-2-Acetolactate: C06010
2-C-Methyl-D-erythritol 2,4-cyclodiphosphate: C11453
2-Acetolactate: C00900
2-Oxoisocaproate: C00233
1-Pyrroline: C15668
4-Aminobutanal: C00555
GABA: C00334
4-Oxobutanoate: C00232
Succinate: C00042
hn: C00205'
L-4-Hydroxy-2-oxoglutarate: C20485
4-Hydroxy-2-oxoglutarate: C01127
2-Keto-3-deoxy-6-phosphogluconate: C04442
Ethanol: C00469
Acetyl adenylate: C05993
L-2-Amino-acetoacetate: C03508
Aminoacetone: C01888
Glycolate: C00160
Tartronate semialdehyde: C01146
Iminoglycine: C15809
Glyoxylate: C00048
Gly: C00037
D-Xylulose: C00310
S-Glutathionyl-L-cysteine: C05526
Putrescine: C00134
L-Arg: C00062
Agmatine: C00179
L-Argininosuccinate: C03406
Citrulline: C00327
L-Ornithine: C00077
delta-1-Pyrroline-5-carboxylate: C04322
2-Dehydropantoate: C00966
Pantoate: C00522
Methanal: C00067
Acaryochloris MBIC11017
Anabaena cylindrica
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Calothrix 336/3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Crinalium epipsammum
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanobium gracile
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Fischerella sp. JSC-11
Geitleriinema PCC 7407
Gloeobacter kilauensis
Gloeobacter violaceus
Gloeocapsa PCC 7428
Halothece PCC 7418
Leptolyngbya PCC 7376
Microcoleus PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc PCC 73102
Nostoc PCC 7107
Nostoc PCC 7524
Nostoc sp. PCC 7120
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 9211
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 0801
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 7335
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-100
Synechococcus KORDI-49
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus PCC 7502
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Synechocystis PCC 6803
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A


0%
25%
50%
75%
100%
S-(Hydroxymethyl)glutathione: C14180
(R)-NADPHX: C20483
(R)-NADHX: C20482
(S)-NADHX: C04856
(S)-NADPHX: C04899
6-Carboxytetrahydropterin: C20239
7,8-Dihydropterin pyrophosphate: C04807
6-Pyruvoyltetrahydropterin: C03684
2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine: C01300
Dihydroneopterin: C04874
Stearoyl-[acyl-carrier protein]: C04088
Dodecanoyl-[acp]: C05223
3-Oxododecanoyl-[acp]: C05756
Butyryl-[acp]: C05745
(R)-3-Hydroxyoctanoyl-[acyl-carrier protein]: C04620
(R)-3-Hydroxybutanoyl-[acyl-carrier protein]: C04618
(R)-3-Hydroxydecanoyl-[acyl-carrier protein]: C04619
Octanoyl-[acp]: C05752
Acetoacetyl-[acp]: C05744
But-2-enoyl-[acyl-carrier protein]: C04246
3-Oxodecanoyl-[acp]: C05753
(2E)-Decenoyl-[acp]: C05754
3-Oxohexanoyl-[acp]: C05746
Hexanoyl-[acp]: C05749
(2E)-Octenoyl-[acp]: C05751
3-Oxooctanoyl-[acp]: C05750
D-3-Hydroxyhexanoyl-[acp]: C05747
(2E)-Hexenoyl-[acp]: C05748
(R)-3-Hydroxypalmitoyl-[acyl-carrier protein]: C04633
HMA: C04688
(2E)-Dodecenoyl-[acp]: C05758
D-3-Hydroxydodecanoyl-[acp]: C05757
Decanoyl-[acp]: C05755
Tetradecanoyl-[acp]: C05761
3-Oxotetradecanoyl-[acp]: C05759
(2E)-Tetradecenoyl-[acp]: C05760
3-Oxohexadecanoyl-[acp]: C05762
Hexadecanoyl-[acp]: C05764
(2E)-Hexadecenoyl-[acp]: C05763
(R)-3-Hydroxystearoyl-[acp]: C16220
(2E)-Octadecenoyl-[acp]: C16221
3-Oxostearoyl-[acp]: C16219
Malonyl-[acp]: C01209
Sedoheptulose 7-phosphate: C05382
beta-D-Glucose 6-phosphate: C01172
D-Mannose 1-phosphate: C00636
beta-D-Fructose 6-phosphate: C05345
Robison ester: C00092
D-Glucosamine phosphate: C00352
D-Mannose 6-phosphate: C00275
6-Phospho-D-glucono-1,5-lactone: C01236
Neuberg ester: C00085
beta-D-Glucose: C00221
alpha-D-Glucose 6-phosphate: C00668
alpha-D-Glucose: C00267
NH3: C00014*
Phosphate: C00009*
Selenide: C01528*
H2O2: C00027*
O2: C00007*
Sulfate: C00059*
H+: C00080*
Sulfite: C00094*
H2: C00282*
H2S: C00283*
Thiosulfate: C00320*
HCl: C01327*
Selenite: C05684*
Fe2+: C14818*
Selenate: C05697*
NH4+: C01342*
H2O: C00001*
Triphosphate: C00536*
PPi: C00013*
Mg2+: C00305*
Co2+: C00175*
Acaryochloris MBIC11017
Anabaena cyl. sp. 90
Anabaena sp. 90
Anabaena variabilis
Arthrospira platensis
Crinalium sp. 3
Calothrix PCC 6303
Calothrix PCC 7507
Chamaesiphon minutus
Chroococcidiopsis thermalis
Cyanobacterium aponinum
Cyanobacterium stanieri
Cyanothece sp. 3
Cyanothece ATCC 51142
Cyanothece PCC 7424
Cyanothece PCC 7425
Cyanothece PCC 7822
Cyanothece PCC 8801
Cyanothece PCC 8802
Cylindrospermum stagnale
Dactylococcopsis salina
Escherichia coli O111:H
Geitlerinema PCC 7407
Globobacter kluyveris
Gloeocapsa sp. PCC 7428
Halothece PCC 7418
Leptolyngbya sp. PCC 7376
Microcoleus sp. PCC 7113
Microcystis aeruginosa
Nodularia spumigena
Nostoc azollae 0708
Nostoc sp. PCC 7107
Nostoc PCC 7107
Nostoc PCC 7120
Nostoc punctiforme
Oscillatoria acuminata
Oscillatoria nigro-viridis
Pleurocapsa PCC 7327
Prochlorococcus AS9601
Prochlorococcus MIT 0601
Prochlorococcus MIT 9215
Prochlorococcus MIT 9301
Prochlorococcus MIT 9303
Prochlorococcus MIT 9312
Prochlorococcus MIT 9313
Prochlorococcus MIT 9515
Prochlorococcus NATL1A
Prochlorococcus NATL2A
Prochlorococcus CCMP1375
Prochlorococcus MED4
Prochlorococcus MIT 0604
Prochlorococcus MIT 9211
Pseudanabaena PCC 7367
Rivularia PCC 7116
Stanieria cyanosphaera
Synechococcus PCC 6301
Synechococcus PCC 7942
Synechococcus CC9311
Synechococcus CC9605
Synechococcus CC9902
Synechococcus JA-2-3Ba(2-13)
Synechococcus JA-3-3Ab
Synechococcus KORDI-49
Synechococcus KORDI-52
Synechococcus PCC 6312
Synechococcus PCC 7002
Synechococcus PCC 7502
Synechococcus RCC307
Synechococcus WH 7803
Synechococcus WH 8109
Thermosynechococcus BP-1
Thermosynechococcus NK55
Trichodesmium erythraeum
Cyanobacterium UCYN-A

Appendix C.

Supplemental material

For this electronic copy of this thesis, the supplemental material is registered under the DOI 10.18452/19215 and available at [http://dx.doi.org/10.18452/19215]. It comprises the following data sets:

Network_reconstructions.zip Zip archive containing the metabolic reconstructions of all 78 bacterial strains as well as the pan-genome. Networks are saved in the universal SBML format. Reactions are labeled by their KEGG reaction IDs and annotated with their physiologically feasible direction, whether they are chemically balanced, and their associated genes.

Table_of_chromosomes_and_plasmids.pdf List of all 78 chromosomes and 136 plasmids considered in this thesis, including the GenBank accession IDs.

Table_of_CLOGs_and_modules.xls Excel sheet enlisting all 58,740 CLOGs and the associated genes. Each CLOG is labeled with the most common annotation as well as EC number and KEGG reaction ID if applicable. CLOGs are grouped into their assigned module and modules are annotated with the strain-specific and the average adjacency score.

Statement of authorship

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected.
I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on March 5th 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

_____

Date / signature of candidate